#2

S&H Form: (2/01)

Attorney Docket No. 1086.1147

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of:

Hidesato MATSUOKA, et al.

Application No.: Unassigned

Group Art Unit: Unassigned

Filed: August 3, 2001

Examiner: Unassigned

For: DOCUMENT ANONYMITY SETTING DEVICE, METHOD AND COMPUTER READABLE RECORDING MEDIUM RECORDING ANONYMITY SETTING PROGRAM

SUBMISSION OF CERTIFIED COPY OF PRIOR FOREIGN APPLICATION IN ACCORDANCE WITH THE REQUIREMENTS OF 37 C.F.R. §1.55

Assistant Commissioner for Patents
Washington, D.C. 2023l

Sir:

In accordance with the provisions of 37 C.F.R. §1.55, the applicant(s) submit(s) herewith a certified copy of the following foreign application:

Japanese Patent Application No. 2001-00641

Filed: January 5, 2001

It is respectfully requested that the applicant(s) be given the benefit of the foreign filing date(s) as evidenced by the certified papers attached hereto, in accordance with the requirements of 35 U.S.C. §119.

Respectfully submitted,
STAAS & HALSEY LLP

Date: August 3, 2001

By: _____
James D. Halsey, Jr.
Registration No. 22,729

700 11th Street, N.W., Ste. 500
Washington, D.C. 20001
(202) 434-1500

日　本　国　特　許　庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出願年月日
Date of Application: ２００１年　１月　５日

出願番号
Application Number: 特願２００１－０００６４１

出願人
Applicant(s): 富士通株式会社

２００１年　５月２５日

特許庁長官
Commissioner,
Japan Patent Office 及川耕造

出証番号　出証特２００１－３０４５１７４

【書類名】　特許願

【整理番号】　0051814

【提出日】　平成13年 1月 5日

【あて先】　特許庁長官殿

【国際特許分類】　G11B 15/20

【発明の名称】　文書匿名化装置、方法及び匿名化プログラムを記録したコンピュータ読取り可能な記録媒体

【請求項の数】　18

【発明者】

【住所又は居所】　神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

【氏名】　松岡　秀達

【発明者】

【住所又は居所】　神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

【氏名】　落谷　亮

【特許出願人】

【識別番号】　000005223

【氏名又は名称】　富士通株式会社

【代理人】

【識別番号】　100079359

【住所又は居所】　東京都港区西新橋３丁目２５番４７号　清水ビル８階

【弁理士】

【氏名又は名称】　竹内　進

【電話番号】　03(3432)1007

【選任した代理人】

【識別番号】　100093584

【住所又は居所】　東京都港区西新橋３丁目２５番４７号　清水ビル８階

【弁理士】

【氏名又は名称】　宮内　　佐一郎

【電話番号】　03(3432)1007

【手数料の表示】

【予納台帳番号】　009287

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9704823

【プルーフの要否】　要

【書類名】　　明細書

【発明の名称】　　文書匿名化装置、方法及び匿名化プログラムを記録したコンピュータ読取り可能な記録媒体

【特許請求の範囲】

【請求項１】

文書を入力する文書入力部と、

前記入力文書から匿名対象表記を抽出し、抽出した匿名対象表記がどの程度の強さで個人を特定できるかを評価する特定度を算出する特定度計算部と、

所定の閾値より大きい特定度を持つ前記入力文書中の表記を匿名化する匿名化処理部と、

を備えたことを特徴とする文書匿名化装置。

【請求項２】

請求項１記載の文書匿名化装置に於いて、

前記特定度計算部は、前記入力文書から人名を抽出し、抽出した人名がどの程度の強さで個人を特定できるかを評価する特定度を算出し、

前記匿名化処理部は、所定の閾値よりも大きい特定度をもつ人名を匿名化することを特徴とする文書匿名化装置。

【請求項３】

請求項１記載の文書匿名化装置に於いて、

前記特定度計算部は、前記入力文書から人名の周辺表記を抽出し、抽出された周辺表記がどの程度の強さで個人を特定できるかを評価する特定度を算出し、

前記匿名化処理部は、所定の閾値よりも大きい特定度をもつ周辺表記を匿名化することを特徴とする文書匿名化装置。

【請求項４】

請求項１記載の文書匿名化装置に於いて、前記特定度算出部は、

入力文書から文を切出す文切出し部と、

切出した文を品詞毎に分解する品詞解析部と、

前記品詞解析結果から人名抽出ルールに基づいて人名を抽出する人名抽出部と統計情報に基づいて抽出した人名の特定度を計算する人名特定度計算部と、

を備えたことを特徴とする文書匿名化装置。

【請求項５】

請求項１記載の文書匿名化装置に於いて、前記特定度算出部は、更に、

前記品詞解析結果から構文解析ルールに基づいて文節間の係り受け関係を示す構文木を作成する構文解析部と、

前記構文解析部で得られた構文木に対し個人特定木抽出ルールに基づいて個人特定木を個人周辺表記として抽出する個人特定木抽出部と、

統計情報に基づいて抽出した個人特定木の特定度を計算する木構造特定度計算部と、

を備えたことを特徴とする文書匿名化装置。

【請求項６】

請求項４又は５記載の文書匿名化装置に於いて、前記特定度算出部は、既存文書に基づいて作成した匿名対象表記、人名か周辺表記かの種別及び特定度を組にした特定度データを登録した基準特定度データベースを備え、前記入力文書から抽出した匿名化表記の計算により求めた特定度を、前記基準匿名度データベースに登録している特定度との重み平均をとって正規化することを特徴とする文書匿名化装置。

【請求項７】

請求項４又は５記載の文書匿名化装置に於いて、前記特定度算出部は、文書データベースの既存文書から文書毎に人名や周辺表記を抽出して特定度を計算し、匿名対象表記、人名又は周辺表記の種類及び特定度の組にした特定度データを登

録した前記基準特定度データベースを作成するデータベース作成部を備えたことをことを特徴とする文書匿名化装置。

【請求項８】

請求項１記載の文書匿名化装置に於いて、更に、前記匿名化処理部で使用する閾値を設定変更する匿名化指示部を設けたことを特徴とする文書匿名化装置。

【請求項９】

請求項１記載の文書匿名化装置に於いて、前記匿名化指示部は、処理文書毎に匿名化処理に使用した閾値を閾値データベースに保存し、新たな入力文書の匿名化処理の際に直前の閾値をデフォルトとして設定することを特徴とする文書匿名化装置。

【請求項１０】

請求項１記載の文書匿名化装置に於いて、前記匿名化処理部は、匿名化不要表記を登録して匿名化不要データベースを持ち、入力文書から抽出された匿名化表記の内、前記匿名化不要データベースに登録されている表記は匿名化しないことを特徴とする文書匿名化装置。

【請求項１１】

請求項１記載の文書匿名化装置に於いて、前記匿名化処理部は、必ず匿名化する表記を登録した匿名化データベースを持ち、該匿名化データベースに登録されている入力文書中の表記は全て匿名化することを特徴とする文書匿名化装置。

【請求項１２】

請求項１記載の文書匿名化装置に於いて、前記匿名化処理部は、入力文書から抽出された匿名化対象表記を伏せ字にすることを特徴とする文書匿名化装置。

【請求項１３】

請求項１記載の文書匿名化装置に於いて、前記匿名化処理部は、入力文書から抽出された匿名化対象表記を、個人を特定しない一般化された表記に置き換えることを特徴とする文書匿名化装置。

【請求項１４】

請求項１記載の文書匿名化装置に於いて、前記匿名化処理部は、入力文書から抽出された匿名化対象表記を、該匿名化対象表記の匿名化に使用する閾値以下の特定度を持つ低特定度表記で置き換えることを特徴とする文書匿名化装置。

【請求項１５】

請求項１記載の文書匿名化装置に於いて、前記匿名化処理部は、入力文書から抽出された匿名化対象表記を暗号化することで匿名化することを特徴とする文書匿名化装置。

【請求項１６】

請求項１６記載の文書匿名化装置に於いて、更に、前記匿名化処理部により暗号化により匿名化された匿名化文書を閲覧する際に、暗号化された匿名化表記を復号化して表示させる復号化指示部を設けたことを特徴とする文書匿名化装置。

【請求項１７】

文書を入力する文書入力ステップと、

前記入力文書から匿名対象表記を抽出し、抽出した匿名対象表記がどの程度の強さで個人を特定かを評価する特定度を算出する特定度計算ステップと、

所定の閾値より大きい特定度を持つ前記入力文書中の表記を匿名化する匿名化処理ステップと、

を備えたことを特徴とする文書匿名化方法。

【請求項１８】

コンピュータに、

文書を入力する文書入力ステップと、

前記入力文書から匿名対象表記を抽出し、抽出した匿名対象表記がどの程度の強さで個人を特定かを評価する特定度を算出する特定度計算ステップと、

所定の閾値より大きい特定度を持つ前記入力文書中の表記を匿名化する匿名化処理ステップと、

を実行させるための匿名化プログラムを記録したコンピュータ読取り可能な記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、文書内の個人を特定するような表現を匿名化する文書匿名化装置、方法及び匿名化プログラムを記録したコンピュータ読取り可能な記録媒体に関し、特に、個人を特定する表現がどの程度の強さで個人を特定できるかを評価して匿名化する文書匿名化装置、方法及び匿名化プログラムを記録したコンピュータ読取り可能な記録媒体に関する。

【０００２】

【従来の技術】

近年、コンピュータを利用したデータ解析の傾向として、顧客からのアンケート回答、苦情、電子メール等の電子化された文書データから事業に役立つ情報を抽出しようとする機運が高まっている。しかし、これらの文書データには個人情報を含まれていることが多く、取扱を間違えると企業の存立に関わる問題となり得る。そこで、文書データを解析する前に、個人情報に関わる情報を適切に隠蔽することが必要となる。

【０００３】

従来、文書データ等に含まれる個人情報は人手により隠蔽化するか、あるいは機械処理が可能な個人名等の直接に個人を特定する表現を隠蔽化する等が行われている。

【０００４】

【発明が解決しようとする課題】

しかしながら、このような従来の個人情報の隠蔽化にあっては、記述されている個人名や個人に関連する周辺表記が、個人情報として保護される情報に属するものか、公的な人物に関する情報のように保護の必要がない情報なのかの区別が作業者にとって判別しづらいため、作業者によって個人情報隠蔽化の適切さが変化するという問題がある。

【０００５】

このため個人情報の隠蔽化を行う作業者の技能と知識は、ある水準を越えている必要があるため、人手による個人情報の隠蔽化は高コストになりやすい。

【０００６】

本発明は、個人情報の隠蔽化を機械化して作業コストを低減し、更に、必要に応じて隠蔽化の度合を調整可能とする文書匿名化装置、方法及び匿名化プログラムを記録したコンピュータ読取り可能な記録媒体を提供することを目的とする。

【０００７】

【課題を解決するための手段】

図１は本発明の概略説明図である。本発明は、文書匿名化装置であり、図１（Ａ）のように、文書を入力する文書入力部１０と、入力文書から匿名対象表記を抽出し、抽出した匿名対象表記がどの程度の強さで個人を特定できるかを評価する特定度を算出する特定度計算部１２と、所定の閾値より大きい特定度を持つ入力文書中の表記を匿名化する匿名化処理部１８とを備えたことを特徴とする。

【０００８】

このため本発明は、文書中の個人を特定するような表現に対して、それがどの程度の強さで個人を特定できるのかを匿名化を行う前に評価しておき、要求される匿名化の水準（閾値）に応じて情報を隠蔽化する。この結果、文書を必要な度合いで自動ないし半自動で匿名化でき、匿名化作業を効率化し作業コストを下げることができる。

【０００９】

ここで特定度計算部１８は、入力文書から人名を抽出し、抽出した人名がどの程度の強さで個人を特定できるかを評価する特定度を算出し、匿名化処理部１８は、所定の閾値よりも大きい特定度をもつ人名を匿名化する。

【００１０】

また特定度計算部１２は、入力文書から人名の周辺表記抽出し、抽出された周辺表記がどの程度の強さで個人を特定できるかを評価する特定度を算出し、このとき匿名化処理部１６は、所定の閾値よりも大きい特定度をもつ周辺表記を匿名化することを特徴とする。ここで周辺表記とは、例えば「大手Ａ社の社長」のように個人名を強く示唆する表記のことである。

【００１１】

特定度算出部１２は、例えば入力文書から文を切出す文切出し部と、切出した文を品詞毎に分解する品詞解析部と、品詞解析結果から人名抽出ルールに基づいて人名を抽出する人名抽出部と、統計情報に基づいて抽出した人名の特定度を計算する人名特定度計算部とを備える。

【００１２】

更に、特定度算出部１２は、品詞解析結果から構文解析ルールに基づいて文節間の係り受け関係を示す構文木を作成する構文解析部と、構文解析部で得られた構文木に対し個人特定木抽出ルールに基づいて個人特定木を周辺表記として抽出する個人特定木抽出部と、統計情報に基づいて抽出した個人特定木（周辺表記）の特定度を計算する木構造特定度計算部とを備える。

【００１３】

特定度算出部１２は、例えば入力文書から文を切出す文切出し部と、切出した文を品詞毎に分解する品詞分解部と、品詞分解部から人名抽出ルールに基づいて人名を抽出する人名抽出部と、品詞解析部から構文解析ルールに基づいて人名に係る文節間の係り受け関係を示す構文木を作成する構文木解析部と、構文解析結果で得られた構文木に対し、個人特定木抽出ルールに基づいて構文特定木を周辺表記として取り出す個人特定木抽出部と、基準文書内での周辺表記と人名の組み合わせを持つ統計情報から、以下の方法で人名や周辺表記が個人を特定する度合

である特定度を計算する特定度計算部を備える。

【００１４】

ここで、周辺表記とは、構文解析の結果、人名と係り受け関係を持つ構文木のことであり、例えば「大手Ａ社の○○社長」には○○社長という人名があって、それに係る修飾句である「大手Ａ社の」が周辺表記である。

【００１５】

特定度算出部１２は、匿名化対象文書に含まれる人名や周辺表記の組み合わせに対して、基準特定度データベース１４中の人名や周辺表記か特定の個人を指す確率を読み出して、基準特定度データベース１４内の全ての個人識別ＩＤについて匿名化対象文書中の人名や周辺表記の組み合わせが持つ個人情報を特定する強さである特定度の計算を行う。基準特定度データベース１４には、個人識別するＩＤと共に人名や周辺表記がその個人を指す確率が登録されている。

【００１６】

個人識別ＩＤがｐである表記数の最大をＮとしたとき、匿名化対象の文書中の人名や周辺表記の組合わせがｐを特定する度合である特定度Ｋ（ｐ）の計算は、ここでは次式で行う。

【００１７】

Ｋ（ｐ）＝（入力文書中の人名又は周辺表記がｐを指す確率の総和）／Ｎ

・・・（１）

ただし、特定度Ｋ（ｐ）の計算方法は、これに限定されるものではなく、入力文書の人名や周辺表記と一致しない人名や周辺表記を持つ個人識別ＩＤについては特定度が低くなり、一致する人名や周辺表記が多い程、表記が表わす個人識別ＩＤの特定度が高くなる性質を持つ計算方法であればよい。

【００１８】

基準特定度データベス一中で全てのｐについて特定度Ｋ（ｐ）を計算しているので、特定度Ｋ（ｐ）を最大にするｐを求めることができ、入力文書や人名や周辺表記の組み合わせはそのｐを越えている可能性が最も高いことになり、最大の特定度Ｋ（ｐ）がある基準値を越えている場合に、入力文書の人名や周辺表記に対し隠蔽化を行うことになる。

【００１９】

ここで特定度の計算に使用する基準特定度データベース１４の作成方法を説明する。基準特定度データベース１４の作成は、データベース作成部１５により行われる。データベース作成部１５は、既存文書の集合である文書データベース７２から文書切出し部によって文書を切り出し、次に文切り出し部によって文単位に分解し、品詞解析部、人名処理部および周辺表記処理部で人名や周辺表記を抽出し、それらがある個人を指す確率を計算し、基準特定度データベース１４に、表記が指す個人を識別するＩＤ、表記の種類、表記、表記が個人を指す確率の４つの組でなる基準特定度データを登録する。

【００２０】

確率計算のためには表記が指す個人を特定する必要があり、そのため電子メールアドレスや住所といった個人を特定する表記を使用する。これらは、以下のような表記の特徴を持っており、その表記の特徴を利用して文書中から取り出す。

【００２１】

（１）電子メールアドレス：abcd@xxx.yyyy.com

（２）住所：○○県○○市○○△丁目△△－△△

これらの個人を特定する表記を個人識別ＩＤに変換する。ある個人識別ＩＤであるｐを使って、ある特定の個人を指す確率Ｐ（ａ→ｐ）や、人名の周辺表記ｓが、特定の個人を指す確率Ｐ（ｓ→ｐ）を以下の式で近似する。文書データベース中で個人を特定する表記を持ち、それがｐを指している文書の集合をＭとすると、

Ｐ（ａ→ｐ）＝（Ｍにおけるあるａの個数）／（Ｍにおける全てのａの数）

・・・（２）

Ｐ（ｓ→ｐ）＝（Ｍにおけるあるａの個数）／（Ｍにおける全てのａの数）

・・・（３）

となる。近似の方法は常識に限定されるものではない。

【００２２】

この計算結果から、個人識別ＩＤ、表記の種類、表記、表記が個人を指す確率の４つの組のデータが基準特定度データベースに基準特定度データとして登録さ

れる。

【００２３】

基準特定度データベース１４から個人を特定する確率を読み出して、特定度計算部で特定度を計算する。計算された特定度を基準値と比較することで、匿名化処理を行うかどうか判別する。

【００２４】

特定度算出部１２は、文書データベースの既存文書から文書毎に人名や周辺表記を抽出して特定度を計算し、匿名対象表記、人名又は周辺表記の種類及び特定度の組にした特定度データを登録した基準特定度データベース１４を作成するデータベース作成部１５を備えるようにしても良い。

【００２５】

本発明の匿名化装置は、更に、匿名化処理部１８で使用する閾値を設定変更する匿名化指示部２０を設ける。このため作業者は、閾値を設定変更しながら匿名化された文書をチェックすることで、個人情報に対する隠蔽化度合を簡単に調整でき、最適な隠蔽化ができる。

【００２６】

匿名化指示部２０は、処理文書毎に匿名化処理に使用した閾値データベース２６に保存し、新たな入力文書の匿名化処理の際に直前の閾値をデフォルトとして設定する。このため検属して文書の匿名化処理をおこなう場合には、一度、最適な閾値の設定調整が済めば、その後は最適化された閾値がデフォルト設定さることで、特に閾値を指示する必要なく処理を進めることができる。

【００２７】

匿名化処理部１８は、匿名化不要表記を登録した匿名化不要データベース２２を持ち、入力文書から抽出された匿名化表記の内、匿名化不要データベースに登録されている表記は匿名化しない。例えば首相や大臣のように公的な人物については、匿名化データベースに登録することで、匿名化の対象から除外する。

【００２８】

匿名化処理部１８は、必ず匿名化する表記を登録した匿名化データベース２４を持ち、この匿名化データベース２６に登録されている入力文書中の表記は全て

匿名化する。例えば企業名、クレジットカードの番号を表す規則、電話番号を表す規則、電子メールのアドレスを表す規則等については、匿名化データベース２６に登録しておくことで、閾値の如何に関わらず確実に匿名化する。

【００２９】

匿名化処理部１８は匿名化処理として次の処理を選択的に行うことができる。

（１）入力文書から抽出された匿名化対象表記を伏せ字にする。

（２）入力文書から抽出された匿名化対象表記を、個人を特定しない一般化された表記に置き換える。

（３）入力文書から抽出された匿名化対象表記を、匿名化対象表記の匿名化に使用する閾値以下の特定度を持つ低特定度表記で置き換える。

（４）入力文書から抽出された匿名化対象表記を暗号化することで匿名化する。

こごて、匿名化処理部１８により暗号化により匿名化された匿名化文書を閲覧する際に、暗号化された匿名化表記を復号化して表示させる復号化指示部３２を設けるようにしても良い。

【００３０】

また本発明は、文書匿名化方法を提供するものであり、図１（Ｂ）のように、

文書を入力する文書入力ステップと；

入力文書から匿名対象表記を抽出し、抽出した匿名対象表記がどの程度の強さで個人を特定できるかを評価する特定度を算出する特定度計算ステップと；

所定の閾値より大きい特定度を持つ入力文書中の表記を匿名化する匿名化処理ステップと；

を備えたことを特徴とする。この文書匿名化方法の詳細は装置構成の場合と同じになる。

【００３１】

更に本発明は、匿名化プログラムを記録したコンピュータ読取り可能な記録媒体を提供するものであり、記録媒体に記録された匿名化プログラムは、

コンピュータに、

文書を入力する文書入力ステップと、

入力文書から匿名対象表記を抽出し、抽出した匿名対象表記がどの程度の強さで

個人を特定できるかを評価する特定度を算出する特定度計算ステップと、
所定の閾値より大きい特定度を持つ入力文書中の表記を匿名化する匿名化処理ステップと、
を実行させる。この記録媒体における匿名化プログラムの詳細も装置構成の場合と同じになる。

【００３２】

【発明の実施の形態】

図２は、本発明による文書匿名化装置の機能構成を示したブロック図であり、コンピュータ装置のプログラム制御により実現される。

【００３３】

図２において、本発明の文書匿名化装置は、文書入力部１０、特定度計算部１２、基準特定度データベース１４、特定度正規化部１６、匿名化処理部１８、匿名化指示部２０、匿名化不要データベース２２、匿名化データベース２４、閾値データベース２６、作業表示部２８及び匿名化文書記憶部３０で構成される。更に匿名化文書記憶部３０に格納された匿名化文書を閲覧するため、復号化指示部３２、判定部３４、閲覧データ作成部３６及び閲覧表示部３８が必要に応じて設けられる。

【００３４】

このような機能構成を持つ本発明の文書匿名化装置につき、各処理部の詳細を説明すると次のようになる。文書入力部１０は匿名化対象文書を入力する。匿名化文書としては、例えはデータ解析対象となる文書が含まれ、例えば顧客からのアンケート回答、苦情、電子メールなどの文書情報を含んでいる。

【００３５】

文書入力部１０で入力した匿名化対象文書は特定度計算部１２に与えられ、特定度が計算される。ここで特定度とは、個人を特定するような表現、即ち人名やその周辺表記に対して、どの程度の強さで個人を特定できるかを評価する値である。

【００３６】

本発明にあっては、特定度算出部１２は、匿名化対象文書から人名や周辺表記を抽出し、抽出した人名や周辺表記が個人を指す確率を文書データベース７２から前記（２）式や（３）式に基づいて算出するか、もしくは、基準特定度データベースから表記が個人を指し示す確率を読み出して特定度を計算し、匿名化状態部１８は、所定の閾値よりも大きい特定度を持つ神明や周辺表記を隠蔽化する。

【００３７】

基準特定度データベース１４には、データベース作成部１５により、後の説明で明らかにするように、十分な量の文書データベース７２を使用して、そこに存在している人名や周辺表記について、前記（２）式や（３）式から表記が個人を指す確率を算出して登録している。

【００３８】

ここで基準特定度データベース１４からは、図９に示すような

（１）個人識別ＩＤ

（２）個人識別ＩＤを示す表記の種別（人名もしくは周辺表記）

（３）個人識別ＩＤを指す表記

（４）表記が個人識別ＩＤを指す確率

の４つを組としたデータが出力される。特定度計算部１２は、基準特定度データベース１４の出力から、（１）式に基づいてある個人識別ＩＤについて特定度を算出する。

【００３９】

図９に示した基準特定度データベース１４の例に従って説明する。この例では前記（１）式で使用するＮは４であったとする。そしてＰ００１の個人を特定する度合である特定度は、（１）式から

（０．＋０．９＋１．０＋０．２）／４＝０．６

となる。Ｐ００３については、人名「松岡」のみが一致したとすると、（１）式から０．２／４＝０．０５　がＰ００３についての特定度となる。全ての個人識別ＩＤについて計算した特定度の中でＰ００１の０．６が最大だったとすると、この０．６と基準値を比較して隠蔽化するかどうかの判定を行う。

【００４０】

特定度算出部１２は、

（１）匿名化表記

（２）特定度

２つの組となる特定度データの形式で匿名化処理部１８に出力する。

【００４１】

匿名化処理部１８は特定度正規化部１６より出力された特定度データを使用して、文書入力部１０より得られた匿名化対象文書について人名や周辺表記を隠蔽化する匿名化処理を行う。

【００４２】

匿名化処理部１８に対しては、匿名化指示部２０より

（１）閾値

（２）使用匿名化方法

（３）処理文書分類

の３つの指示値が与えられる。

【００４３】

匿名化指示部２０による閾値は、作業者がキーボードやマウスなどの入力デバイスを使用して設定変更できる値であり、人名閾値と周辺表記を対象とした個人特定木閾値を個別に設定することができる。この匿名化指示部２０から設定された閾値は特定度正規化部１６から得られた特定度データの特定度と比較され、閾値以上の特定度をもつ人名及び周辺表記について隠蔽するための匿名化処理が行われる。

【００４４】

匿名化指示部２０による閾値の設定は、直接閾値の数値を入力させる方法以外に、特定度に対応したスライドバーをマウスで操作する方法、閾値の設定ウィンドウを開いてウィンドウ項目の中から閾値を選択する方法など、適宜の視覚的な操作を含む。

【００４５】

匿名化指示部２０による使用匿名化方法の指示に対応して、匿名化処理部１８には次の匿名化方法が設けられている。

（１）伏せ字化

（２）一般化

（３）低特定度化

（４）暗号化

まず伏せ字化は、匿名化対象とする人名や周辺表記を伏せ字に使用する記号を選択し、選択した記号で匿名化対象を全て置き換える。例えば「佐藤」といった人名を「××」とする。

【００４６】

一般化は匿名化対象中の固有名詞を一般的な表記で置き換える。このため、一般的な表記で置き換えるための一般化ルールを匿名化処理部１８は備えている。この一般化ルールには例えば

ルール１：人名は「Ａ」に置き換える。

【００４７】

ルール２：企業名は「Ａ」に置き換える。

などが記述されている。

【００４８】

低特定度化は、匿名化の対象をより特定度の低い表記で置き換える。この低特定度化のため、人名と特定度の組及び個人特定木と特定度の組について、小さい特定度のものを基準特定度データベース１４から検索し、このとき匿名化対象となっている特定度より低い特定度の検索した表記を用いて匿名化対象を置き換える。具体的には、人名の場合には基準特定度データベース１４から小さい特定度を持つ人名を検索し、検索結果で匿名化対象の人名を置き換えることで匿名化する。

【００４９】

また周辺表記となる個人特定木の場合には、基準特定度データベース１４から匿名化対象の個人特定木と共通する分節を含む個人特定木で特定木が小さい個人特定木を検索し、この検索した個人特定木で匿名化対象を置き換える。もし基準特定度データベース１４から検索した結果が匿名化対象の特定度以下でない場合には、低特定度化による匿名化はできないことから、処理の失敗を作業者に知ら

せることになる。更に、暗号化は、匿名化の対象を所定の暗号規則に従って暗号化する。

【００５０】

匿名化処理部１８に対しては、匿名化不要データベース２２と匿名化データベース２４が設けられている。匿名化不要データベース２２には匿名化を行う表記や識別するための規則が登録されている。匿名化不要データベース２２に登録されている例としては例えば次のものがある。

（１）首相、大臣などの公人の人名

（２）芸能人の人名

（３）首相、大臣のような、公の人物であることを示す周辺表記を持つ人物を識別する規則

このため、匿名化処理部１８で文書入力部１０より入力した匿名化対象文書を匿名化する際に、匿名化不要データベース２２を参照し、そこに登録している人名や表記については匿名化を一切行わないことになる。

【００５１】

匿名化データベース２４には匿名化処理の際に必ず匿名化を行う表記や、これを識別するための規則が登録されている。例えば匿名化データベース２４には次の表記や規則が登録されている。

（１）企業名

（２）クレジットカードの番号を表わす規則

（３）電話番号を表す規則

（４）電子メールのアドレスを表わす規則

このため匿名化処理部１８にあっては、文書入力部１０から入力した匿名化対象文書の中に匿名化データベース２４に登録している表記や識別規則に該当する表記がある場合には、特定度計算部１２及び特定度正規化部１６で求められた特定度の如何に関わらず強制的に匿名化を行って認定することになる。

【００５２】

匿名化処理部１８で匿名化された文書は作業表示部２８に表示され、作業者は匿名化の結果を確認しながら匿名化指示部２０により閾値や使用匿名化方法を変

更し、必要とする隠蔽化が行われた匿名化文書を作成することができる。匿名化処理部１８で作成された匿名化文書は匿名化文書記憶部３０に保存される。　匿名化文書記憶部３０のレコード形式としては、文書コードなどを使用した処理文書分類、匿名化処理情報、匿名化文書の形式をもって保存する。もちろん指示情報には、文書匿名化処理の際に匿名化指示部２４で指示された閾値使用匿名化方法が含まれている。

【００５３】

閾値データベース２６には匿名化文書記憶部３０に記憶された匿名化文書レコードより得た閾値と匿名化方法が処理文書分類である分類コードによって登録されている。特に閾値データベース２６の先頭位置には、最新の匿名化文書に関する閾値により匿名化方法が格納されており、匿名化処理部１８にあっては、この閾値データベース２６の先頭位置の閾値及び匿名化方法を匿名化指示部２０によるデフォルトの設定内容としている。

【００５４】

このため作業者にあっては、匿名化指示部２０により閾値や使用匿名化方法の指示を行わなくても、直前に行われた匿名化文書における閾値及び匿名化方法が自動的に匿名化処理部１８に設定されることになる。

【００５５】

匿名化文書記憶部３０に保存されている匿名化文書は、閲覧データ作成部３６により読み出して閲覧表示部３８に表示して閲覧することができる。このうち暗号化による匿名化文書については、復号化指示部３２からの暗号化方法に対応したパスワードの入力で匿名化文書中の暗号化部分を元の人名や周辺表記に復号化して閲覧することができる。

【００５６】

復号化指示部３２からのパスワードは判定部３４で判定され、パスワードに対応した復号化方法が閲覧データ作成部３６に指示され、暗号化された表記を復号して閲覧することができる。この匿名化処理部１８における暗号化と閲覧時の復号化については、後の説明で更に明らかにされる。

【００５７】

図３は、図２における本発明の文書匿名化処理のフローチャートである。図３において、ステップＳ１で匿名化処理要求の有無をチェックしており、作業者による匿名化処理要求を判別すると、ステップＳ２に進み、文書入力部１０より匿名化対象文書を入力する。続いてステップＳ３で特定度計算部１２により匿名化対象文書に含まれる人名やその周辺表記である匿名化対象について特定度を計算する。具体的には、基準特定度データベース１４の参照で個人名の指す確率を取得する。続いてステップＳ４で特定度に基づいて匿名化処理を行う。

【００５８】

ステップＳ４で匿名化処理が済むと、ステップＳ５で匿名化文書を保存する。続いてステップＳ６で閲覧要求をチェックしており、閲覧要求があればステップＳ７に進み、保存している匿名化文書の閲覧データを作成して表示する。そしてステップＳ８で終了指示があれば、一連の処理を終了する。

【００５９】

図４は、図２の特定度計算部１２の詳細を示した機能構成のブロック図である。特定度計算部１２は、文切出部４０、品詞解析部４２、人名処理部４４及び周辺表記処理部４６で構成される。人名処理部４４には人名抽出部４８、人名特定度計算部５０及び人名抽出ルール５２が設けられている。

【００６０】

また周辺表記処理部４６には構文解析部５４、個人特定木抽出部５６、木構造特定度計算部５８、構文解析ルール６０及び個人特定木抽出ルール６２が設けられている。

【００６１】

図２の文書入力部１０で入力された匿名化文書は、図４の特定度計算部１２における文切出部４０に与えられ、文単位に分解して切り出した文を品詞解析部４２に入力する。品詞解析部４２は形態素解析などを利用して切り出した文を品詞情報付きの品詞に分解し、人名処理部４４と周辺表記処理部４６のそれぞれに出力して人名処理及び周辺表記処理をそれぞれ独立に行わせる。

【００６２】

まず人名処理部４４を説明すると、品詞解析部４２から文を品詞ごとに分解し

て受けた人名抽出部４８は、人名抽出ルール５２を用いて人名を抽出し、人名特定度計算部５０に出力する。人名抽出ルール５２としては「ｉｆ～ｔｈｅｎ～」形式によって次の規則が登録されている。

規則５２１：　ｉｆ［姓］，［名］ｔｈｅｎ人名として抽出

規則５２２：　ｉｆ［姓］ｔｈｅｎ人名として抽出

規則５２３：　ｉｆ［名］ｔｈｅｎ人名として抽出

この「ｉｆ～ｔｈｅｎ」の規則において、ｉｆの次の条件部では品詞名を［］で表わす。また、この条件部で「，」で繋がった品詞は連続しているものを表わしている。

【００６３】

このような人名抽出ルール６０の規則５２１，５２２，５２３によって、規則に一致する品詞パターンを持った文字列として人名が抽出される。例えば規則５２１により連続した姓名から人名が抽出される。また規則５２２により姓から人名が抽出される。更に規則５２３により名から人名が抽出される。

【００６４】

人名特定度計算部５０は、匿名化対象となる表記により、基準特定度データベース１４を参照し、表記が個人を指す確率を取得する。

【００６５】

次に周辺表記処理部４６に説明する。周辺表記処理部４６の構文解析部５６は、品詞解析部４２から得られた品詞ごとに分解した文を対象に、構文解析を利用して分節間の係り受け関係を示す木構造、即ち構文木を作成する。このとき構文解析部５４は構文解析ルール６０を使用する。構文解析ルール６０には「ｉｆ～ｔｈｅｎ」形式で次の規則が記述されている。

規則６０１：　ｉｆ［名詞句］，［助詞「の」］，［人名］ｔｈｅｎ
［人名名詞句］（［名詞句］→＜修飾＞→［人名］）

規則６０２：　ｉｆ［名詞句］，［助詞「の」］，［人名名詞句］ｔｈｅｎ
［人名名詞句］（［名詞句］→＜修飾＞→［人名名詞句］）

この規則６０１，６０２において、ｉｆの後ろの条件部は品詞間に複数の要素が入っている条件を表している。またｔｈｅｎの直後には、条件部が成立した場合

にひとまとめにした品詞を記述し、（　）内に生成する要素間の関係を記述する。更に＜　＞の中には生成される関係に付けられた名前を表わしている。

【００６６】

この構文解析による構文木の生成を具体的に説明すると次のようになる。いま次のような文があったとする。

「△△社の社長でピアニストの○○は××ホールで演奏した」

規則６０１は名詞句と名詞句の間に助詞の「の」が入っているときに全体を名詞句とし、

［名詞句］→＜修飾＞→［人名］

の修飾関係を生成し、これは例文の「ピアニストの」が「○○」を修飾している木構造に対応している。したがって、この場合の木構造として図６が得られる。

【００６７】

図６のように得られた構文木に対し、次の個人特定木抽出部５６は、個人特定木抽出ルール６２を適用して、個人を特定する部分木を個人特定木として抽出し、木構造特定度計算部５８に出力する。

【００６８】

個人特定木抽出ルール６２には次のような規則が登録されている。

規則６２１：　ｉｆ［名詞句］→＜修飾＞→＜人名＞ｔｈｅｎ

個人特定木として抽出

規則６２２：　ｉｆ［名詞句］→＜修飾＞→［人名名詞句］ｔｈｅｎ

個人特定木として抽出

即ち規則６２１は、人名などを修飾する名詞句を個人特定木として抽出することに対応する。例えば「ピアニスト」が「○○」を修飾している木構造から

「ピアニストの」→＜修飾＞→「○○」

を個人特定木として抽出することができる。この例では、これ以外に図７のような木構造がそれぞれ個人特定木として抽出される。

【００６９】

木構造特定度計算部５８は人名特定度計算部５０と同様、基準特定度データベース１４の参照により、木構造が個人を指す確率を取得し、特定度を計算する。

【００７０】

図７は、図４の特定度計算部１２における処理のフローチャートである。この特定度計算処理にあっては、ステップＳ１で匿名化対象文書から文を切り出し、ステップＳ２で品詞ごとに分解する品詞解析を行い、人名処理及び周辺表記処理のそれぞれに供給する。

【００７１】

人名処理にあっては、ステップＳ３で人名抽出を行い、ステップＳ４で人名特定度を計算し、併せて基準特定度データベース１４の参照で得られた特定度と共に出力する。また周辺表記処理にあっては、ステップＳ５で構文解析を行った後、ステップＳ６で個人特定木抽出処理を行い、ステップＳ７で木構造特定度計算を行うと共に、基準特定度データベース１４から基準特定度を取得し、正規化処理に出力する。

【００７２】

図８は、図１の特定度計算部１２に設けているデータベース作成部１５の機能を取り出している。このデータベース作成部１５は、文書データベース７２に格納されている十分な量の文書を対象に基準特定度データベース１４を作成する。

【００７３】

このためデータベース作成部１５にあっては、文書データベース７２から対象文書を切り出す文切出部４０が設けられ、切り出した文書は文切出部４０に与えられる。

【００７４】

データベース作成部１５の文切出部４０と品詞解析部４２は、図４の特定度算出部１２のブロックのものと同一のものが使用される。周辺表記処理部４６－１は、個人を特定する周辺表記である電子メールアドレスや住所等を抽出し、それを個人識別ＩＤに置き換える。

【００７５】

電子メールアドレスや住所は、以下のような表記の特徴をもっており、この表記を使って本分から取り出す。

【００７６】

電子メールアドレス：abcd@xxx.yyyy.com

住所：○○県○○市○○△丁目△△－△△

人名処理部４４－１及び周辺表記処理部４６－１では、図４の人名処理部４４及び周辺表記処理部４６と同じ機構で人名や周辺表記を抽出する。抽出した人名や周辺表記については、前記（２）式や（３）式に従って、人名や周辺表記が個人を指す確率が計算される。

【００７７】

そして人名処理部４４－１および周辺表記処理部４６－１で作成された、個人識別ＩＤ、表記の種類、表記、及び表記が個人を指す確率の４つの組となる特定度データは、例えば図９のように、基準特定度データベース１４に基準特定度データとして登録される。

【００７８】

データベース作成部１５の処理の流れを図１０に示す。既存文書を厚めた文書データベース７２から文書を切り出し、その文書を文に分解して人名と周辺表記である個人特定木を取り出すところまでは、図５の特定度計算部と同様である。周辺表記の中で個人を特定する電子メールアドレスや住所等の表記の特徴から判別して個人識別ＩＤを作成し、個人識別ＩＤ毎に、人名や周辺表記が個人を指す確率を（２）式もしくは（３）式から計算し。図９のような４つの組のデータとして基準特定度データベースに登録する。

【００７９】

図９は、基準特定度データベース１４の登録内容の例であり、種別、表記、基準特定度の項目によって基準特定度データが登録され、種別としては人名及び周辺表記を表す構文木が格納されている。

【００８０】

図１０は、図８のデータベース作成部１５の処理のフローチャートである。この基準特定度データベースの作成処理にあっては、ステップＳ１で文書データベース７２から文書を切り出して標準文書を作成し、ステップＳ３で品詞ごとに分解する品詞解析を行う。

【００８１】

この品詞解析結果はステップＳ４，Ｓ５の人名処理及びステップＳ６～Ｓ９の周辺表記処理のそれぞれに与えられ、独立に人名抽出と人名特定度の計算、及び構文解析、個人特定木抽出に基づく木構造特定度、個人識別ＩＤの作成が行われる。そしてステップＳ１０で最終的に、基準特定度データベース１４に図９のように基準特定度データを登録する。

【００８２】

この基準特定度データベースの作成処理は、本発明の匿名化装置を使用する前の準備段階で基本的に行うが、運用中においても必要に応じて適宜に文書データベース７２を更新して、新たな文書データを対象に基準特定度データベース１４の再構築を行うことが望ましい。

【００８３】

図１１は、図３のステップＳ５における匿名化処理の詳細のフローチャートである。この匿名化処理にあっては、ステップＳ１で匿名化指示部２０からの指示に基づき、匿名化処理部１８で使用する匿名化情報を決定する。匿名化条件は匿名化指示部２０からの指示がないときは、閾値データベース２６に基づいて行う。

【００８４】

図１２は、図２の閾値データベース２６の登録内容であり、処理文書分類となる分類コード、閾値及び匿名化方法の項目で構成されている。この内、処理文書分類の分類コード００となる先頭位置には、直前の匿名化処理で使用した直前閾値とその匿名化方法が登録されている。この分類コード００の先頭位置の閾値及び匿名化方法は、図１１の匿名化処理における匿名化指示のデフォルト条件として設定される。

【００８５】

このためステップＳ１の匿名化条件決定の際に匿名化指示部２０による作業者の指示がなければ、図１２の閾値データベース２６の先頭位置となる分類コード００の匿名化方法、この場合には「伏せ字化」と「直前閾値」が匿名化条件として設定される。

【００８６】

ステップＳ２で匿名化条件が承認されると、ステップＳ３で匿名化条件を決定した条件に変更する。そしてステップＳ４で匿名化対象文書について匿名化表記である人名や周辺表記を検索し、ステップＳ５で匿名化表記があれば、ステップＳ６で匿名化不要データベース２２の参照により匿名化不要表記を検索する。

【００８７】

ステップＳ７で匿名化不要表記があれば、それ以降の処理をスキップする。匿名化不要表記がなければ、ステップＳ８で匿名化表記について求められている特定度を匿名化条件として設定した閾値と比較し、閾値以上であればステップＳ９の置換処理に入る。

【００８８】

この置換処理は、伏せ字化、一般化、低特定度化、暗号化のいずれかの処理となる。そしてステップＳ１０で全ての匿名化表記検索が終了したか否かチェックし、終了していなければ再びステップＳ４に戻り、同様な処理を繰り返し、全ての匿名化表記の処理が済めば一連の処理を終了する。

【００８９】

図１３は、図１１のステップＳ９における置換処理の詳細のフローチャートである。図１３において、まずステップＳ１で匿名化条件として伏せ字化の指示の有無をチェックしており、伏せ字化の指示であればステップＳ２に進み、予め準備された伏せ字に使用する記号を選択し、ステップＳ３で閾値以上の特定度を持つ匿名化対象表記を対象に伏せ字への置き換えを行う。

【００９０】

一方、ステップＳ４で匿名化条件として一般化の指示が判別された場合には、予め準備している一般化ルールを参照して、ステップＳ５で一般表記を選択し、ステップＳ６で閾値以上の特定度を持っている匿名化対象表記について、選択した一般表記への置き換えを行う。

【００９１】

またステップＳ７で匿名化条件として低匿名化の指示を判別した場合には、ステップＳ１０に進み、基準特定度データベース１４より小さい特定度を持つ人名または周辺表記としての個人特定木をステップＳ１０で検索する。

【００９２】

このデータベース検索に対し、ステップＳ１１で低特定度表記があれば、ステップＳ１２で検索した特定度表記への置き換えを行う。一方、ステップＳ１１で低特定度表記がデータベースから検索できなかった場合には、ステップＳ１４で作業者に対し失敗を通知して処理を終了する。

【００９３】

一方、ステップＳ７で低特定度化でなかった場合には、この場合は暗号化であることからステップＳ８に進み、暗号化表記を生成し、ステップＳ９で匿名化対象表記を暗号化表記に置き換える。

【００９４】

そしてステップＳ３，ステップＳ６，ステップＳ９またはステップＳ１２のいずれかの置き換えが済むと、ステップＳ１３で匿名化文書と表記データを出力し、必要があれば再度、匿名化条件の設定を行って匿名化処理を繰り返し、匿名化終了であれば匿名化文書記憶部３０に匿名化文書を保存するようになる。

【００９５】

ここでステップＳ８，Ｓ９における復号化による匿名化処理を説明すると次のようになる。暗号化表記による置き換えとしては、例えば暗号化によって匿名化した箇所の開始位置に暗号化したことを示すコード＜ＣＲＹＰＴ＞を埋め込み、終了位置に暗号化の範囲が終了したことを示すコード＜／ＣＲＹＰＴ＞を埋め込む。また復号化方法を示す場合には開始コード＜ＣＲＹＰＴ＞の部分を＜ＣＲＹＰＴ　ＭＥＴＨＯＤ＝”復号化方法”＞として復号化方法を記述する。

【００９６】

例えば「△△さんはプログラムの解析を行った」を暗号化により匿名化すると、次のようになる。

「<CRYPT METHOD="METHOD1">%abc$12DE;KsrBX </CRYPT>さんはプログラムの解析を行った。」

この暗号化は匿名化対象文書の「△△」を暗号化した結果が「%abc$12DE;KsrBX 」、復号化の方法が「ＭＥＴＨＯＤ１」の場合に、匿名化対象の表記「△△」を暗号化表記で置き換えたものである。

【００９７】

この例では「ＭＥＴＨＯＤ１」で指定された復号化方法の中にパスワードや公開範囲を指定しておき、暗号化後のデータが外部に流出したとしても、匿名化対象「△△」という人名は復号化しない限り読み取ることができないようにする。

【００９８】

また暗号化と復号化の方法を何通りか用意しておき、それぞれの方法ごとに対応するパスワードを変えておくことで、暗号化した表記ごとに復号化される部分とそうでない部分とを区別できるようにし、閲覧者ごとに読取り可能な範囲を変化させることもできる。更に暗号化の方法を記述する方法として、暗号化された匿名化文書に暗号化部分を示す情報を埋め込む方法以外に、暗号化した表記の位置情報や暗号化方法を匿名化文書、本文とは別文書で記憶させてもよい。

【００９９】

このような暗号化による置換処理で得られた匿名化文書については、図２の復号化指示部３２、判定部３４、閲覧データ作成部３６に示すように、暗号化方法と復号化方法に対応して定められたパスワードを使用した復号化指示部３２からの指示を判定部３４に対し行うことで、パスワードに基づいた復号化方法より匿名化文書記憶部３０に格納されている暗号化表記で置換した匿名化文書を読み出し、暗号化表記の部分を元の人名や周辺表記に復号して閲覧表示部３８で見ることができる。

【０１００】

図１４は、図２の作業表示部２８に表示された匿名化作業画面８８であり、文書入力部１０より入力された匿名化対象文書として電子メール９０が表示されている。この匿名化作業画面８８の右側には匿名化条件を設定するウィンドウ９２が設けられ、ウィンドウを開くことで原文９２－１が表示されていることを示している。

【０１０１】

このような原文の匿名化作業画面８８について、図１５のようにウィンドウ９２を開いて、その選択内容から閾値として低レベル９２－２を指示し、この状態で実行キー９４をマウスクリックすると、閾値レベルを低レベルとした本発明に

よる文書匿名化処理が実行され、匿名化文書９６の表示が行われる。

【０１０２】

この低レベルの閾値を設定した匿名化文書９６を図１４の原文である電子メール９０と対比すると、企業名「情報媒体」、所属名「情報機器」、名「英達」が、それぞれ「○○○○」、「××××」「△△」に置換されている。また原文の電子メール９０におけるメールアドレス、電話番号、ファックス番号及び住所についても、それぞれ匿名化の表記での置換が行われている。

【０１０３】

図１６は、ウィンドウ９２の閾値レベルを高レベル９２－３に設定した場合の実行キー９４のマウスクリックによる処理結果としての匿名化文書９６を表示した匿名化作業画面８８である。

【０１０４】

このように閾値レベルを高レベルとした場合には、図１５の閾値レベルを低レベルとしたい場合には匿名化されていなかった人名「佐藤」「松岡」についても、「▽▽」「△△」のように匿名化表記への置換が行われ、個人情報に対する隠蔽度が更に高められる。

【０１０５】

次に本発明による文書匿名化プログラムを記録したコンピュータ読取り可能な記録媒体の実施形態を説明する。本発明による文書匿名化プログラムは、図３のフローチャートに示した処理ステップを備えている。

【０１０６】

即ち本発明の記録媒体に格納された匿名化プログラムは、コンピュータに文書を入力する文書入力ステップと、入力文書から匿名化対象表記を抽出し、抽出した匿名化対象表記がどの程度の強さで個人を特定するかを評価する特定度を参照する特定度計算ステップと、所定の閾値より大きい特定度を持つ入力文書中の表記を匿名化する匿名化処理ステップとを実行させる。

【０１０７】

この記憶媒体には、ＣＤ－ＲＯＭやフロッピィディスクなどのリムーバブルな可搬型記録媒体、回線によりプログラムを提供するプログラム提供者の記憶装置

、更にはプログラムをインストールした処理装置のＲＡＭやハードディスクなどのメモリ装置がある。また記録媒体によって提供された文書匿名化プログラムは処理装置にローディングされ、その主メモリ上で実行される。

【０１０８】

また本発明により提供される記録媒体に格納された文書匿名化プログラムは、図２における文書入力部１０、特定度計算部１２、基準特定度データベース１４、特定度正規化部１６、匿名化処理部１８、匿名化指示部２０、匿名化不要データベース２２、匿名化データベース２４及び閾値データベース２６、更に匿名化文書記憶部３０の処理機能を備えればよい。

【０１０９】

なお上記の実施形態にあっては、図２のように特定度計算部１２で匿名化対象表記について特定度を計算と基準特定度データベース１４の両方から求め、特定度正規化部１６で正規化する場合を例にとっているが、特定度計算部１２で匿名化表記について基準特定度データベース１４から取得し、基準特定度データベース１４にない場合に計算により特定度を求めるようにしてもよい。この場合には計算または基準特定度データベース１４のいずれかから特定度が求まることから、特定度正規化部１６による正規化は行わない。

【０１１０】

また本発明は、その目的と利点を損なわない適宜の変形を含む。更に本発明は上記の実施形態に示した数値による限定は受けない。

【０１１１】

（付記）

（付記１）

文書を入力する文書入力部と、

前記入力文書から匿名対象表記を抽出し、抽出した匿名対象表記がどの程度の強さで個人を特定できるかを評価する特定度を算出する特定度計算部と、

所定の閾値より大きい特定度を持つ前記入力文書中の表記を匿名化する匿名化処理部と、

を備えたことを特徴とする文書匿名化装置。（１）

（付記２）

付記１記載の文書匿名化装置に於いて、

前記特定度計算部は、前記入力文書から人名を抽出し、抽出した人名がどの程度の強さで個人を特定できるかを評価する特定度を算出し、

前記匿名化処理部は、所定の閾値よりも大きい特定度をもつ人名を匿名化することを特徴とする文書匿名化装置。（２）

（付記３）

付記１記載の文書匿名化装置に於いて、

前記特定度計算部は、前記入力文書から人名の周辺表記を抽出し、抽出された周辺表記がどの程度の強さで個人を特定できるかを評価する特定度を算出し、

前記匿名化処理部は、所定の閾値よりも大きい特定度をもつ周辺表記を匿名化することを特徴とする文書匿名化装置。（３）

（付記４）

付記１記載の文書匿名化装置に於いて、前記特定度算出部は、

入力文書から文を切出す文切出し部と、

切出した文を品詞毎に分解する品詞解析部と、

前記品詞解析結果から人名抽出ルールに基づいて人名を抽出する人名抽出部と、

統計情報に基づいて抽出した人名の特定度を計算する人名特定度計算部と、

を備えたことを特徴とする文書匿名化装置。（４）

（付記５）

付記１記載の文書匿名化装置に於いて、前記特定度算出部は、更に、

前記品詞解析結果から構文解析ルールに基づいて文節間の係り受け関係を示す構文木を作成する構文解析部と、

前記構文解析部で得られた構文木に対し個人特定木抽出ルールに基づいて個人特定木を個人周辺表記として抽出する個人特定木抽出部と、

統計情報に基づいて抽出した個人特定木の特定度を計算する木構造特定度計算部と、

を備えたことを特徴とする文書匿名化装置。（５）

（付記６）

付記４又は５記載の文書匿名化装置に於いて、前記特定度算出部は、既存文書に基づいて作成した匿名対象表記、人名か周辺表記かの種別及び特定度を組にした特定度データを登録した基準特定度データベースを備え、前記入力文書から抽出した匿名化表記の計算により求めた特定度を、前記基準匿名度データベースに登録している特定度との重み平均をとって正規化することを特徴とする文書匿名化装置。（６）

（付記７）

付記４又は５記載の文書匿名化装置に於いて、前記特定度算出部は、文書データベースの既存文書から文書毎に人名や周辺表記を抽出して特定度を計算し、匿名対象表記、人名又は周辺表記の種類及び特定度の組にした特定度データを登録した前記基準特定度データベースを作成するデータベース作成部を備えたことをことを特徴とする文書匿名化装置。（７）

（付記８）

付記１記載の文書匿名化装置に於いて、更に、前記匿名化処理部で使用する閾値を設定変更する匿名化指示部を設けたことを特徴とする文書匿名化装置。（８）

（付記９）

付記１記載の文書匿名化装置に於いて、前記匿名化指示部は、処理文書毎に匿名化処理に使用した閾値を閾値データベースに保存し、新たな入力文書の匿名化処理の際に直前の閾値をデフォルトとして設定することを特徴とする文書匿名化装置。（９）

（付記１０）

付記１記載の文書匿名化装置に於いて、前記匿名化処理部は、匿名化不要表記を登録して匿名化不要データベースを持ち、入力文書から抽出された匿名化表記の内、前記匿名化不要データベースに登録されている表記は匿名化しないことを特徴とする文書匿名化装置。（１０）

（付記１１）

付記１記載の文書匿名化装置に於いて、前記匿名化処理部は、必ず匿名化する表記を登録した匿名化データベースを持ち、該匿名化データベースに登録されている入力文書中の表記は全て匿名化することを特徴とする文書匿名化装置。（１１

）

（付記１２）

付記１記載の文書匿名化装置に於いて、前記匿名化処理部は、入力文書から抽出された匿名化対象表記を伏せ字にすることを特徴とする文書匿名化装置。（１２）

（付記１３）

付記１記載の文書匿名化装置に於いて、前記匿名化処理部は、入力文書から抽出された匿名化対象表記を、個人を特定しない一般化された表記に置き換えることを特徴とする文書匿名化装置。（１３）

（付記１４）

付記１記載の文書匿名化装置に於いて、前記匿名化処理部は、入力文書から抽出された匿名化対象表記を、該匿名化対象表記の匿名化に使用する閾値以下の特定度を持つ低特定度表記で置き換えることを特徴とする文書匿名化装置。（１４）

（付記１５）

付記１記載の文書匿名化装置に於いて、前記匿名化処理部は、入力文書から抽出された匿名化対象表記を暗号化することで匿名化することを特徴とする文書匿名化装置。（１５）

（付記１６）

付記１５記載の文書匿名化装置に於いて、更に、前記匿名化処理部により暗号化により匿名化された匿名化文書を閲覧する際に、暗号化された匿名化表記を復号化して表示させる復号化指示部を設けたことを特徴とする文書匿名化装置。（１６）

（付記１７）

文書を入力する文書入力ステップと、

前記入力文書から匿名対象表記を抽出し、抽出した匿名対象表記がどの程度の強さで個人を特定できるかを評価する特定度を算出する特定度計算ステップと、

所定の閾値より大きい特定度を持つ前記入力文書中の表記を匿名化する匿名化処理ステップと、

を備えたことを特徴とする文書匿名化方法。（１７）

（付記１８）

付記１７記載の文書匿名化方法に於いて、

前記特定度計算ステップは、前記入力文書から人名を抽出し、抽出した人名がどの程度の強さで個人を特定できるかを評価する特定度を算出し、

前記匿名化処理ステップは、所定の閾値よりも大きい特定度をもつ人名を匿名化することを特徴とする文書匿名化方法。

【０１１２】

（付記１９）

付記１７記載の文書匿名化方法に於いて、

前記特定度計算ステップは、前記入力文書から人名の周辺表記抽出し、抽出された周辺表記がどの程度の強さで個人を特定できるかを評価する特定度を算出し、

前記匿名化処理ステップは、所定の閾値よりも大きい特定度をもつ周辺表記を匿名化することを特徴とする文書匿名化方法。

【０１１３】

（付記２０）

付記１７記載の文書匿名化方法に於いて、前記特定度算出ステップは、

入力文書から文を切出す文切出しステップと、

切出した文を品詞毎に分解する品詞解析ステップと、

前記品詞解析結果から人名抽出ルールに基づいて人名を抽出する人名抽出ステップと、

統計情報に基づいて抽出した人名の特定度を計算する人名特定度計算ステップと、

を備えたことを特徴とする文書匿名化方法。

【０１１４】

（付記２１）

付記１７記載の文書匿名化方法に於いて、前記特定度算出ステップは、更に、前記品詞解析結果から構文解析ルールに基づいて文節間の係り受け関係を示す構文木を作成する構文解析ステップと。

前記構文解析ステップで得られた構文木に対し個人特定木抽出ルールに基づいて

個人特定木を個人周辺表記として抽出する個人特定木抽出ステップと、
統計情報に基づいて抽出した個人特定木の特定度を計算する木構造特定度計算ステップと、
を備えたことを特徴とする文書匿名化方法。

【０１１５】

（付記２２）

付記２０又は２１記載の文書匿名化方法に於いて、前記特定度算出ステップは、前記入力文書から抽出した匿名化表記の計算により求めた特定度を、既存文書に基づいて作成した匿名対象表記、人名か周辺表記かの種別及び特定度を組にした特定度データを、登録した基準特定度データベースに登録している特定度との麻績み平均をとって正規化することを特徴とする文書匿名化方法。

【０１１６】

（付記２３）

付記２０又は２１記載の文書匿名化方法に於いて、前記特定度算出ステップは、文書データベースの既存文書から文書毎に人名や周辺表記を抽出して特定度を計算し、匿名対象表記、人名又は周辺表記の種類及び特定度の組にした特定度データを登録した前記基準特定度データベースを作成するデータベース作成ステップを備えたことをことを特徴とする文書匿名化方法。

【０１１７】

（付記２４）

付記１７記載の文書匿名化方法に於いて、更に、前記匿名化処理ステップで使用する閾値を設定変更する匿名化指示ステップを設けたことを特徴とする文書匿名化方法。

【０１１８】

（付記２５）

付記１７記載の文書匿名化方法に於いて、前記匿名化指示ステップは、処理文書毎に匿名化処理に使用した閾値を閾値データベースに保存し、新たな入力文書の匿名化処理の際に直前の閾値をデフォルトとして設定することを特徴とする文書匿名化方法。

【０１１９】

（付記２６）

付記１７記載の文書匿名化方法に於いて、前記匿名化処理ステップは、入力文書から抽出された匿名化表記の内、匿名化不要データベースを参照して登録されている表記は匿名化しないことを特徴とする文書匿名化方法。

【０１２０】

（付記２７）

付記１７記載の文書匿名化方法に於いて、前記匿名化処理ステップは、必ず匿名化する表記を登録した匿名化データベースに登録されている入力文書中の表記は全て匿名化することを特徴とする文書匿名化方法。

【０１２１】

（付記２８）

付記１７記載の文書匿名化方法に於いて、前記匿名化処理ステップは、入力文書から抽出された匿名化対象表記を伏せ字にすることを特徴とする文書匿名化方法。

【０１２２】

（付記２９）

付記１７記載の文書匿名化方法に於いて、前記匿名化処理ステップは、入力文書から抽出された匿名化対象表記を、個人を特定しない一般化された表記に置き換えることを特徴とする文書匿名化方法。

【０１２３】

（付記３０）

付記１７記載の文書匿名化方法に於いて、前記匿名化処理ステップは、入力文書から抽出された匿名化対象表記を、該匿名化対象表記の匿名化に使用する閾値以下の特定度を持つ低特定度表記で置き換えることを特徴とする文書匿名化方法。

【０１２４】

（付記３１）

付記１７記載の文書匿名化方法に於いて、前記匿名化処理ステップは、入力文書から抽出された匿名化対象表記を暗号化することで匿名化することを特徴とする

文書匿名化方法。

【０１２５】

（付記３２）

付記３２記載の文書匿名化方法に於いて、更に、前記匿名化処理ステップにより暗号化により匿名化された匿名化文書を閲覧する際に、暗号化された匿名化表記を復号化して表示させる復号化指示ステップを設けたことを特徴とする文書匿名化方法。

【０１２６】

（付記３３）

コンピュータに、

文書を入力する文書入力ステップと、

前記入力文書から匿名対象表記を抽出し、抽出した匿名対象表記がどの程度の強さで個人を特定できるかを評価する特定度を算出する特定度計算ステップと、

所定の閾値より大きい特定度を持つ前記入力文書中の表記を匿名化する匿名化処理ステップと、

を実行させるための匿名化プログラムを記録したコンピュータ読取り可能な記録媒体。（１８）

【０１２７】

【発明の効果】

以上説明してきたように本発明によれば、文書中の個人を特定するような表現に対し、それがどの程度の強さで個人を特定できるのかを、匿名化を行う前に評価しつつ、要求される匿名化の水準（閾値）に応じて対象となる表記を匿名化して適切に隠蔽化でき、これによって文書を必要な度合いで自動ないし半自動で匿名化でき、匿名化作業を効率化し、作業コストを大幅に低減することができる。

【図面の簡単な説明】

【図１】

本発明の概略説明図

【図２】

本発明の機能構成のブロック図

【図３】

本発明による文書匿名化処理のフローチャート

【図４】

図２の特定度計算部の機能構成のブロック図

【図５】

構文解析で得られた周辺表記の構文木の説明図

【図６】

図５の構文木から抽出された個人特定木の説明図

【図７】

図４の特定度計算処理のフローチャート

【図８】

図２の特定度計算部に設けているデータベース作成部の機能構成のブロック図

【図９】

基準特定度データベースの説明図

【図１０】

図８の基準特定度データベース作成処理のフローチャート

【図１１】

本発明における匿名化処理のフローチャート

【図１２】

図２の閾値データベースの説明図

【図１３】

図１１の置換処理のフローチャート

【図１４】

本発明で処理する原文作業画面の説明図

【図１５】

低レベルの閾値を指示した場合の本発明による匿名化文書の画面説明図

【図１６】

高レベルの閾値を指示した場合の本発明による匿名化文書の画面説明図

【符号の説明】

１０：文書入力部

１２：特定度計算部

１４：基準特定度データベース

１５：データベース作成部

１８：匿名化処理部

２０：匿名化指示部

２２：匿名化不要データベース

２４：匿名化データベース

２６：閾値データベース

２８：作業表示部

３０：匿名化文書記憶部

３２：復号化指示部

３４：判定部

３６：閲覧データ作成部

３８：閲覧表示部

４０：文切出部

４２：品詞解析部

４４：人名処理部

４４－１：人名処理部（データベース作成部内）

４６：周辺表記処理部

４６－１：周辺表記処理部（データベース作成部内）

４８：人名抽出部

５０：人名抽出ルール

５２：人名特定度計算部

５４：構文解析部

５６：個人特定木抽出部（周辺表記抽出部）

５８：木構造特定度計算部（周辺表記特定度計算部）

６０：構文解析ルール

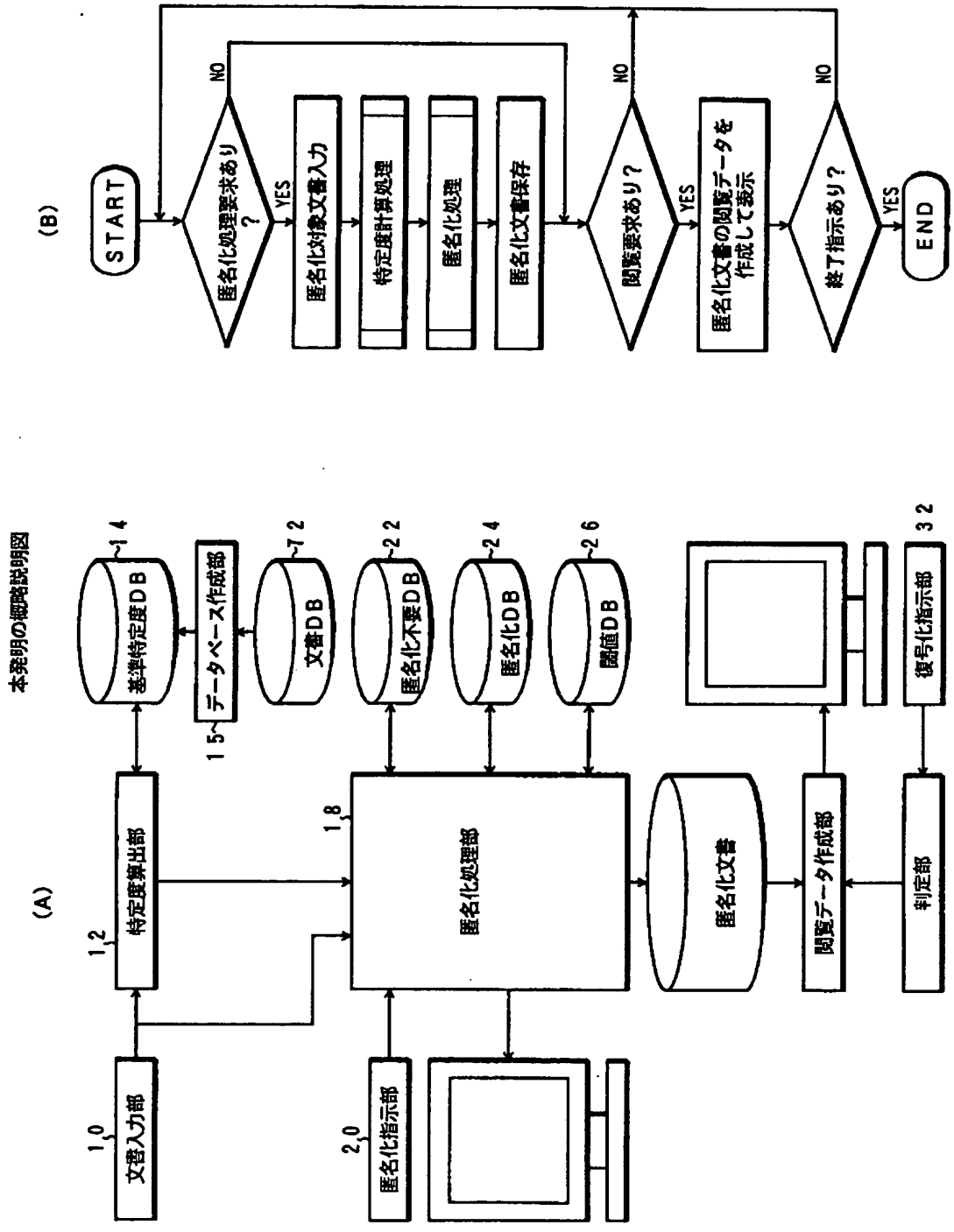
６２，８６：個人特定木抽出ルール

７２：文書データベース

７４：文書切出部

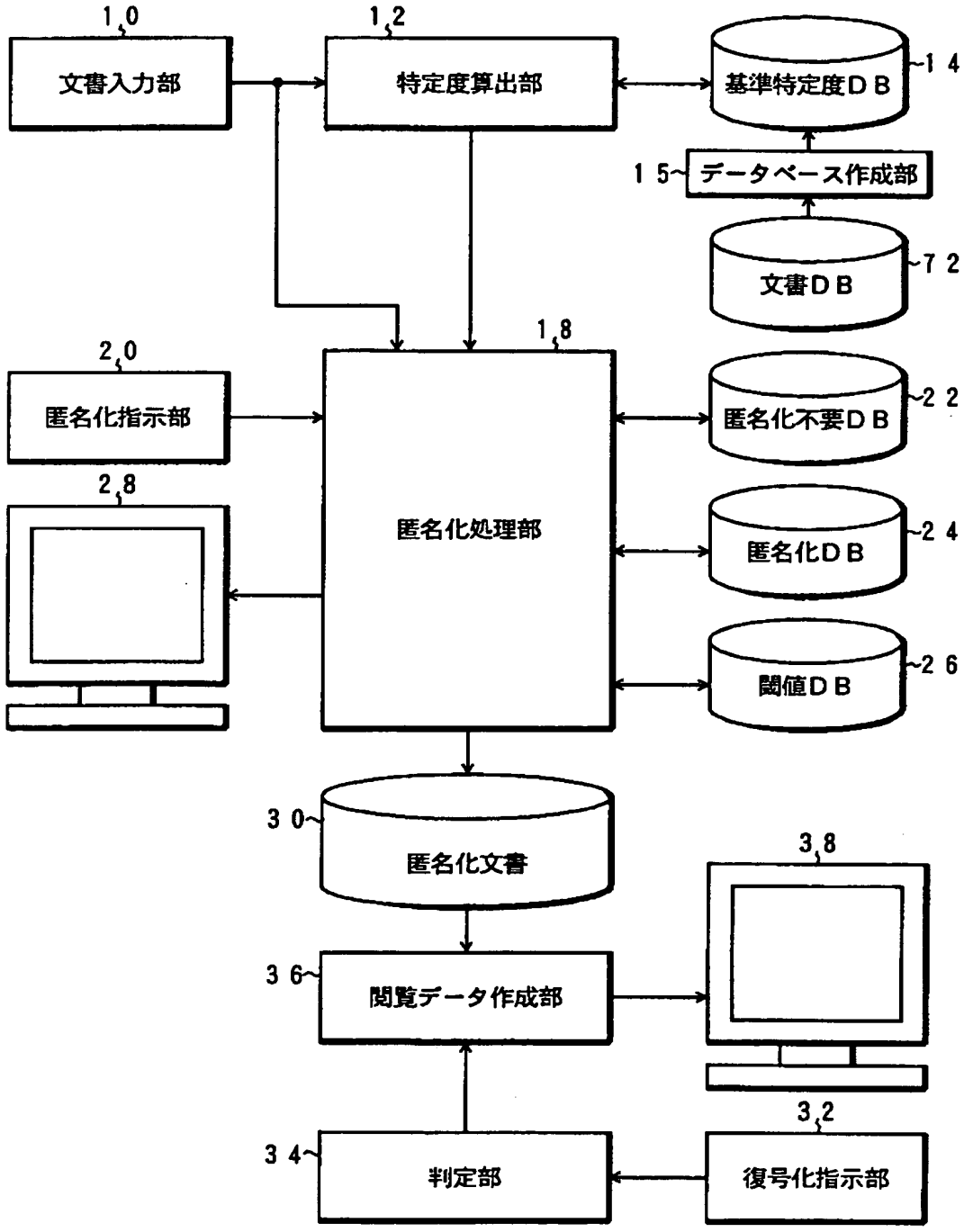
【書類名】　　図面

【図１】


(B)
START
匿名化処理要求あり？
NO
YES
匿名化対象文書入力
特定度計算処理
匿名化処理
匿名化文書保存
閲覧要求あり？
NO
YES
匿名化文書の閲覧データを作成して表示
終了指示あり？
NO
YES
END
本発明の概略説明図
(A)
基準特定度ＤＢ
14
データベース作成部
15
文書ＤＢ
72
匿名化不要ＤＢ
22
匿名化ＤＢ
24
閾値ＤＢ
26
復号化指示部
32
特定度算出部
12
匿名化処理部
18
匿名化文書
閲覧データ作成部
判定部
文書入力部
10
匿名化指示部
20

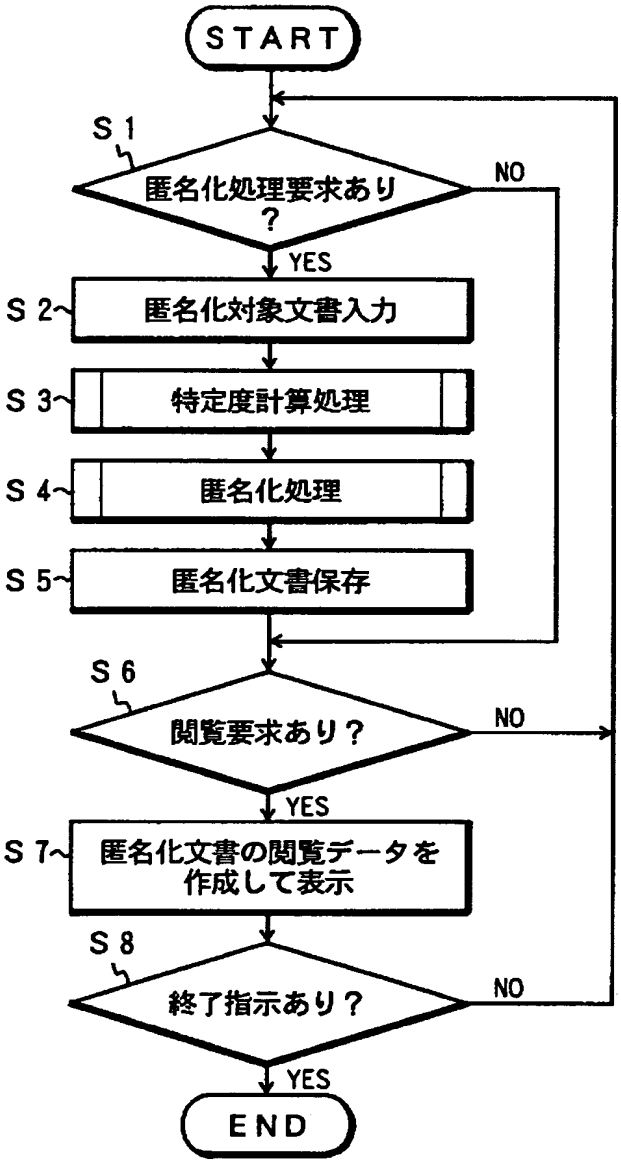
【図２】

本発明の機能構成のブロック図



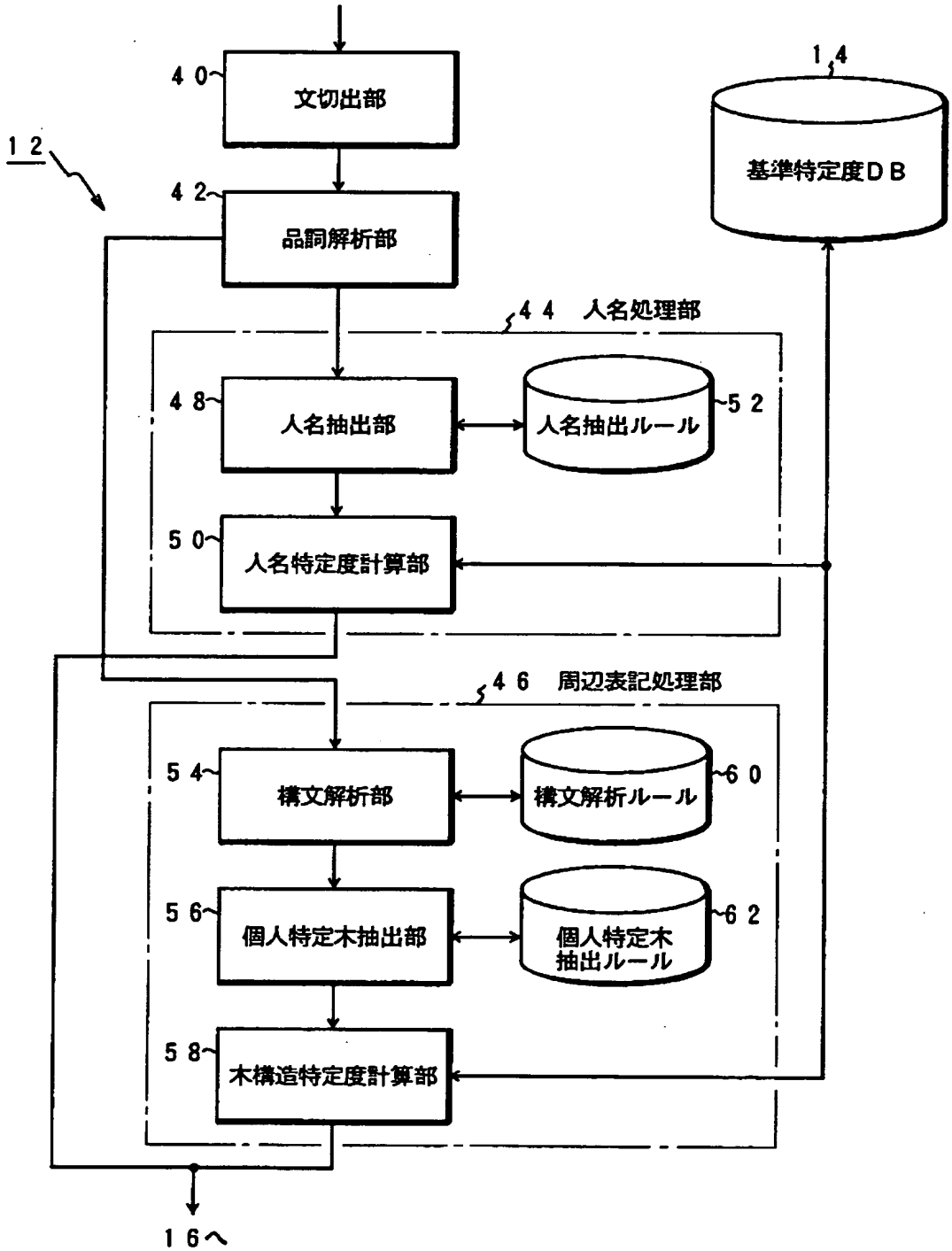
【図３】

本発明による文書匿名化処理のフローチャート



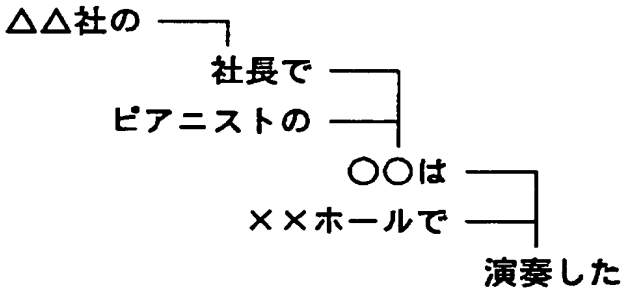
【図４】

図２の特定度計算部の機能構成のブロック図



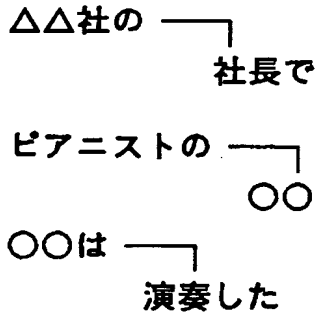
【図５】

構文解析で得られた周辺表記の構文木の説明図



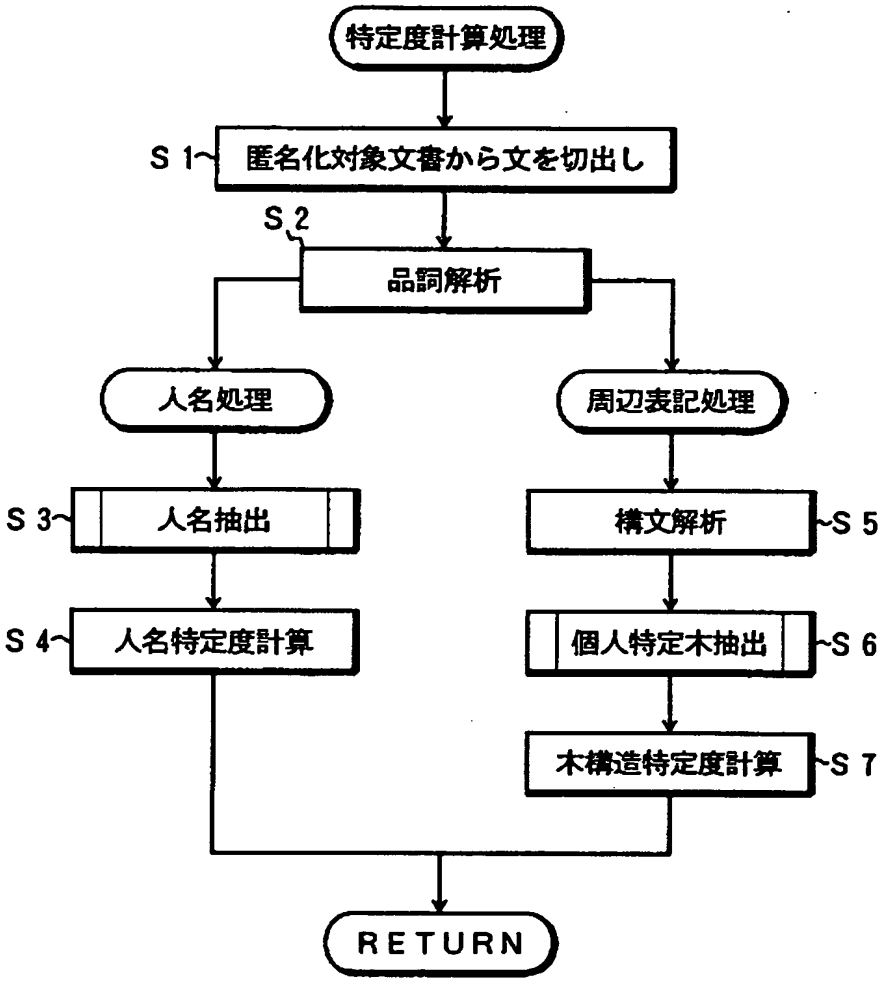
【図６】

図５の構文木から抽出された個人特定木の説明図



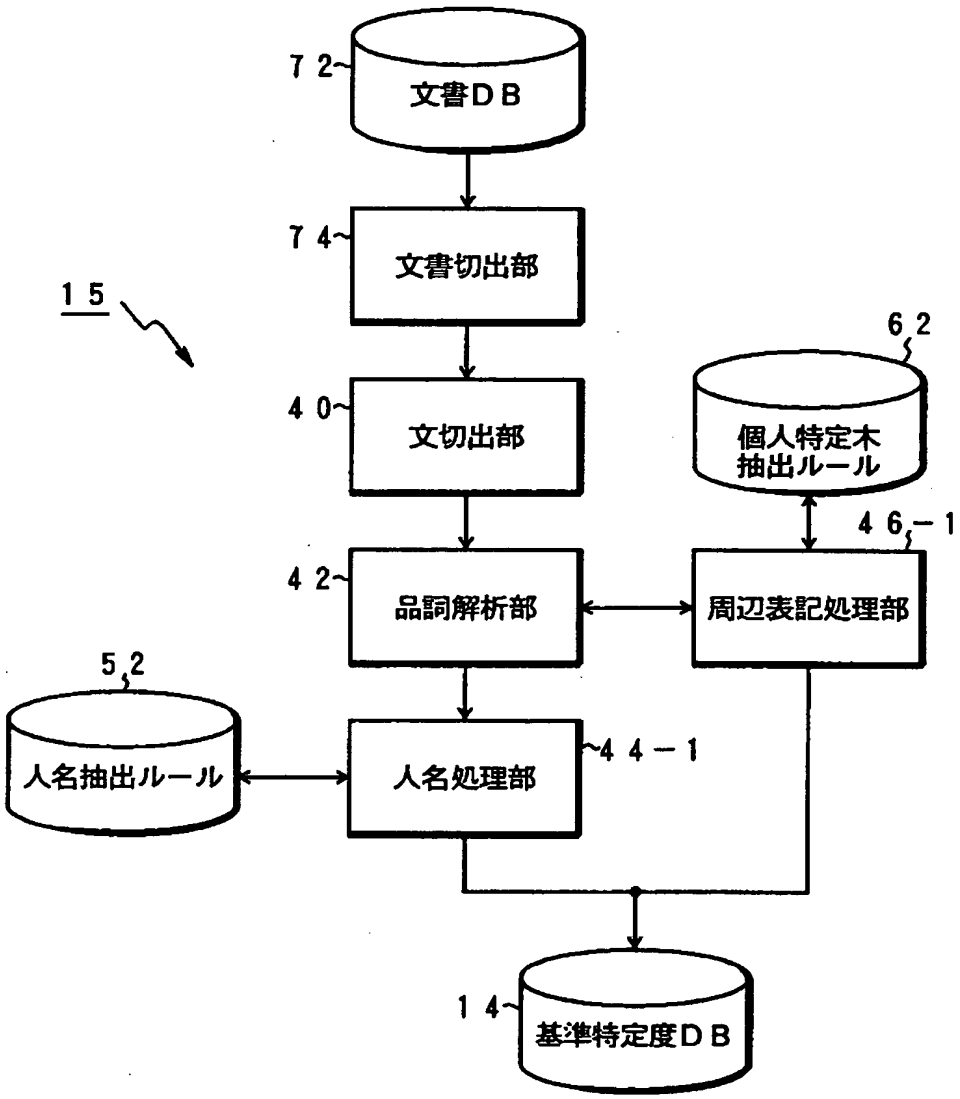
【図７】

図４の特定度計算処理のフローチャート



【図８】

図２の特定度計算部に設けているデータベース作成部の機能構成のブロック図



【図９】

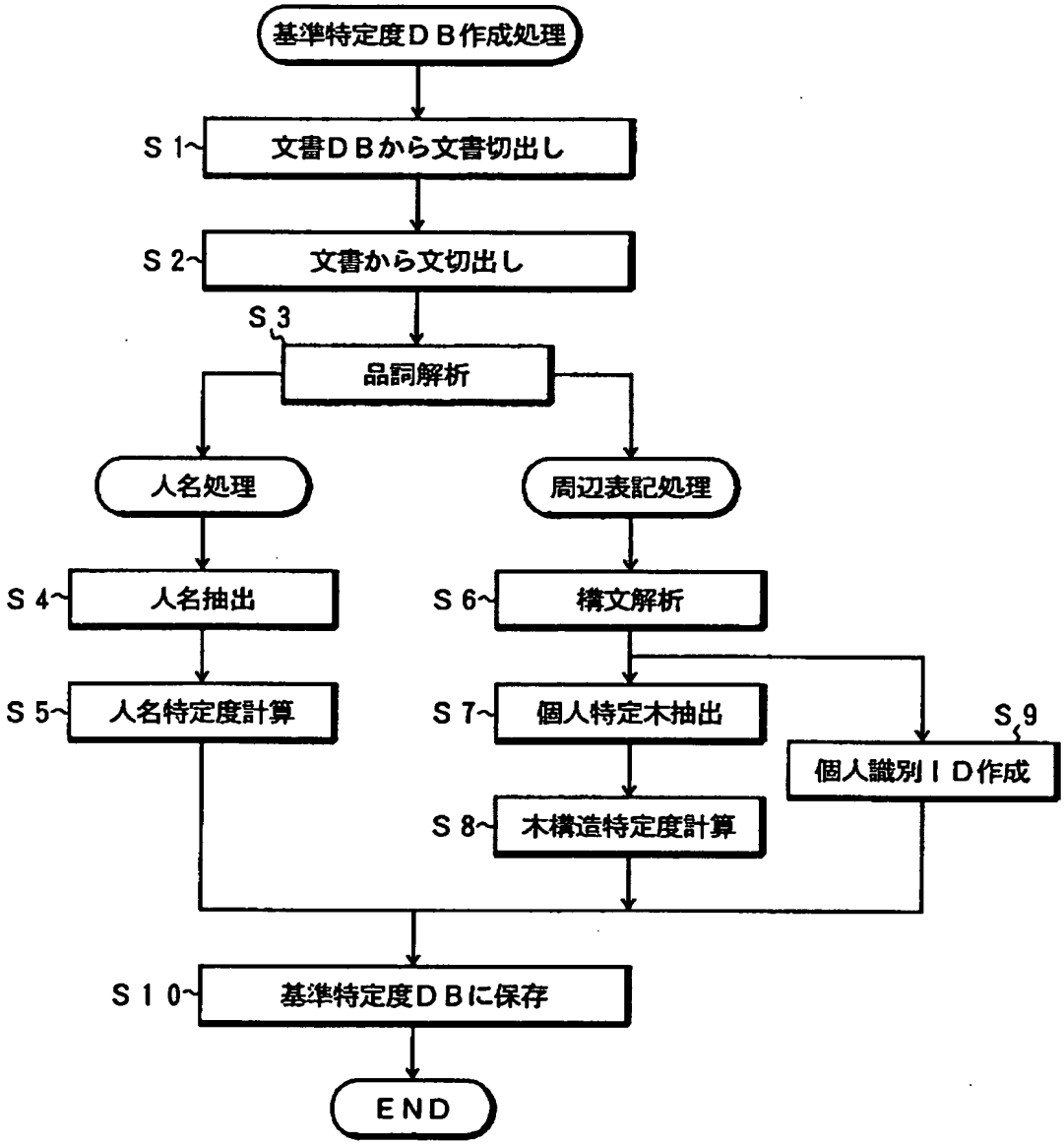
図２の閾値データベースの説明図

14

個人識別ＩＤ	表記の種類	表記	確率
Ｐ００１	人名	松岡	0.3
Ｐ００１	人名	松岡英達	0.9
Ｐ００１	周辺	h-matsu@karino.kaisha.co.jp	1.0
Ｐ００１	周辺	情報機器営業部の	0.2
……	……	……	…
Ｐ００３	人名	松岡	0.2
Ｐ００３	周辺	埼玉県草加市	0.6
……	……	……	…
Ｐ００９	人名	埼玉県草加市	0.6
……	……	……	…

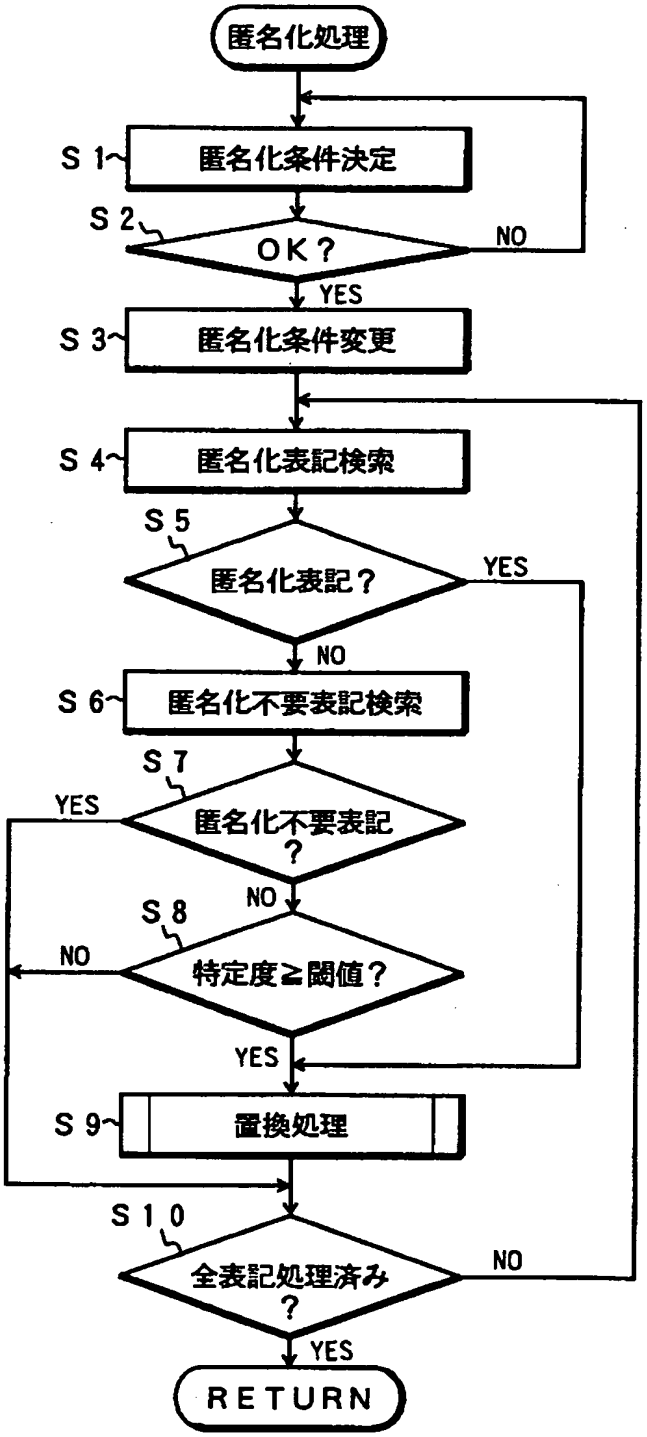
【図１０】

図８の基準特定度データベース作成処理のフローチャート



【図１１】

本発明における匿名化処理のフローチャート



【図１２】

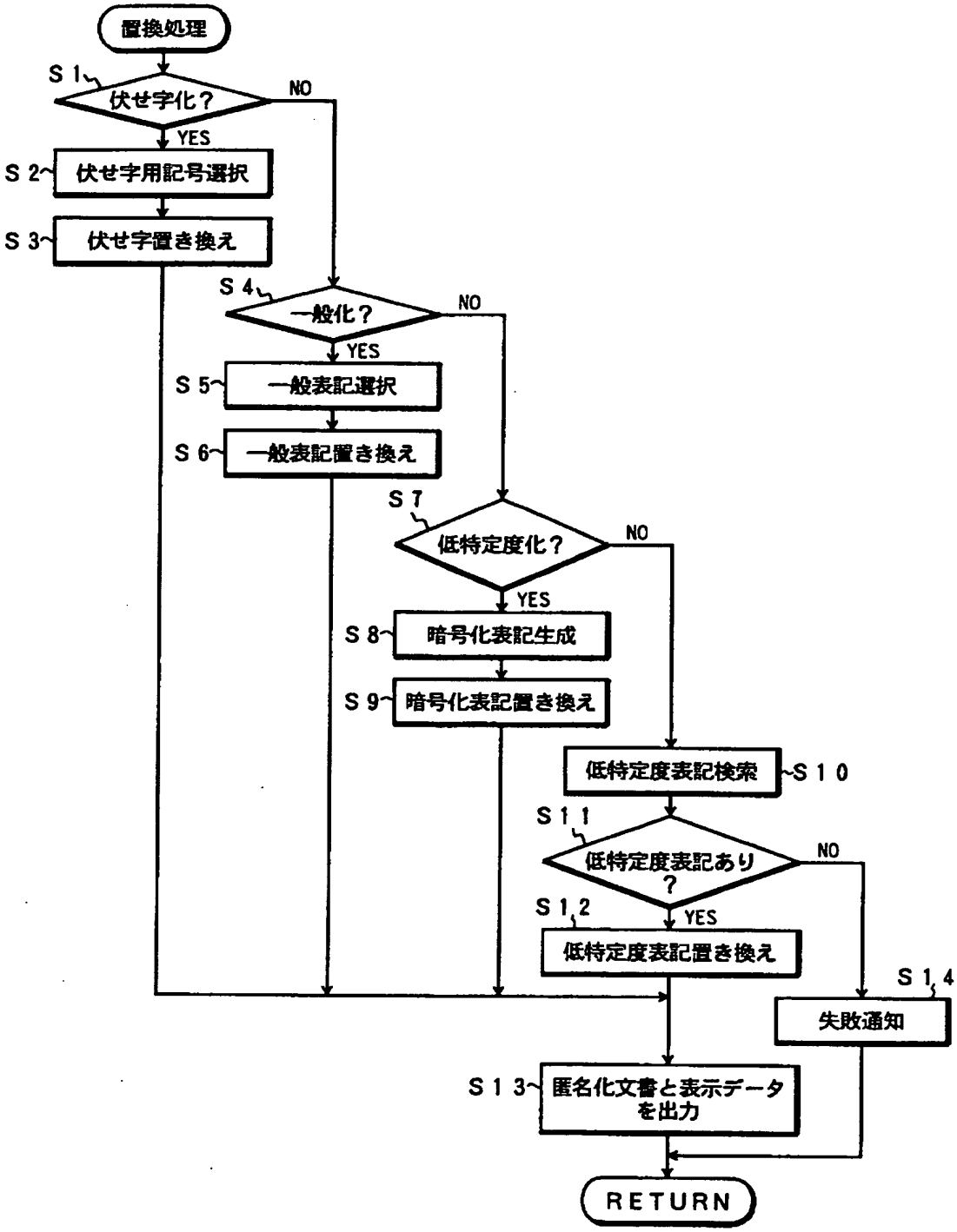
図２の閾値データベースの説明図

26

処理文書分類	閾　値	匿名化方法
００	直前閾値	伏せ字化
０１	ＴＨ１	一般化
０２	ＴＨ２	一般化
０３	ＴＨ３	低特定度化
０４	ＴＨ４	暗号化
・ ・ ・	・ ・ ・	・ ・ ・
ｎ	ＴＨｎ	伏せ字化

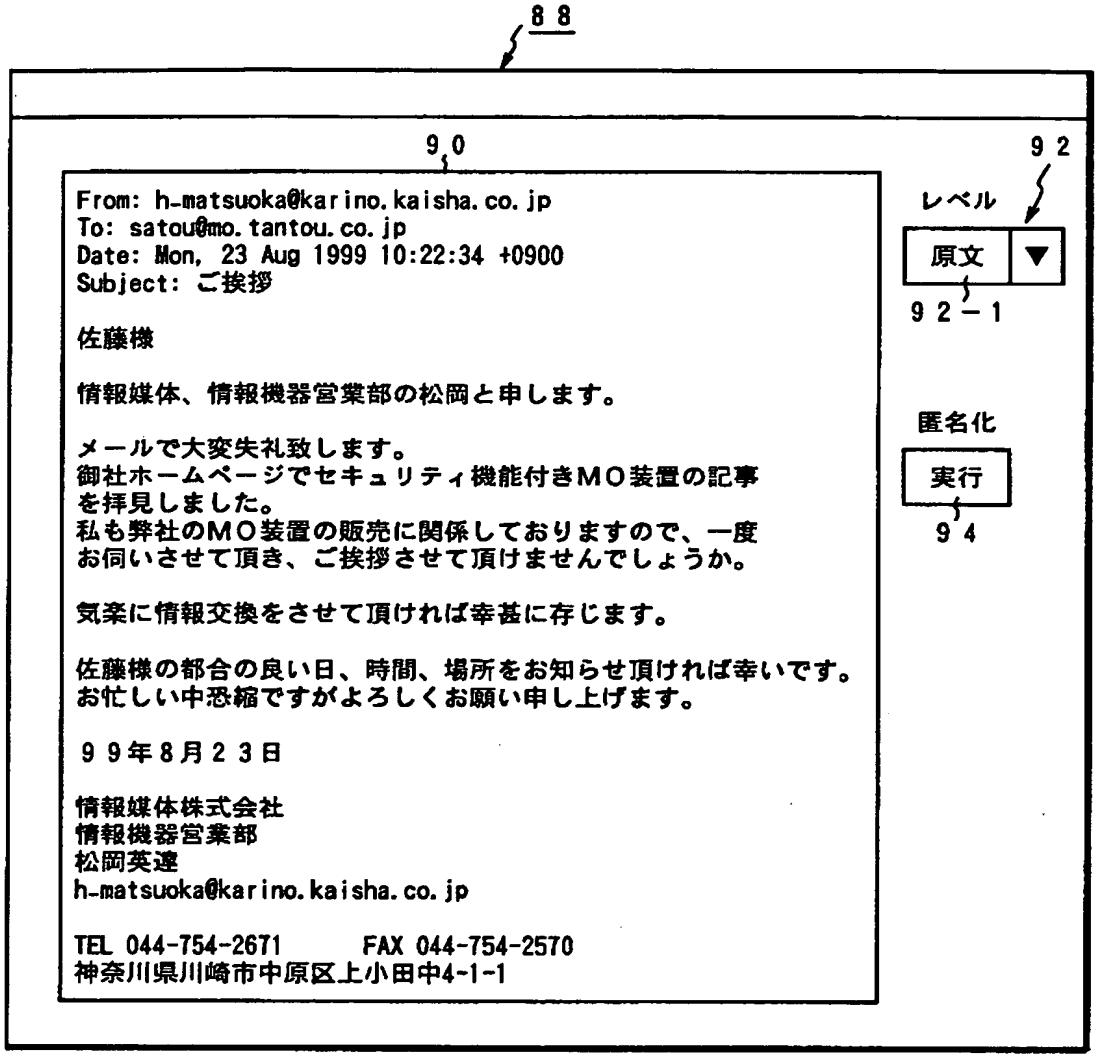
【図１３】

図１１の置換処理のフローチャート



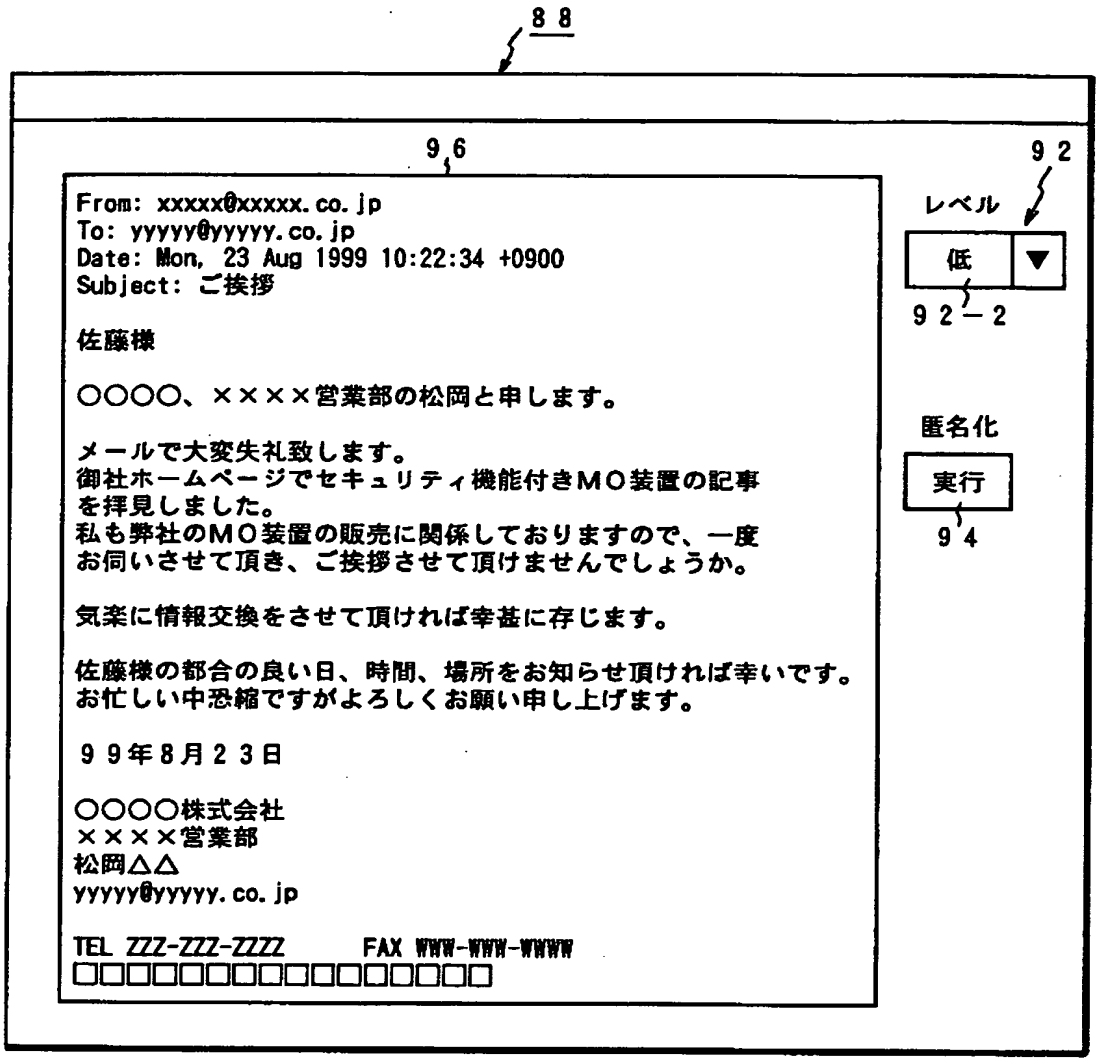
【図１４】

本発明で処理する原文作業画面の説明図



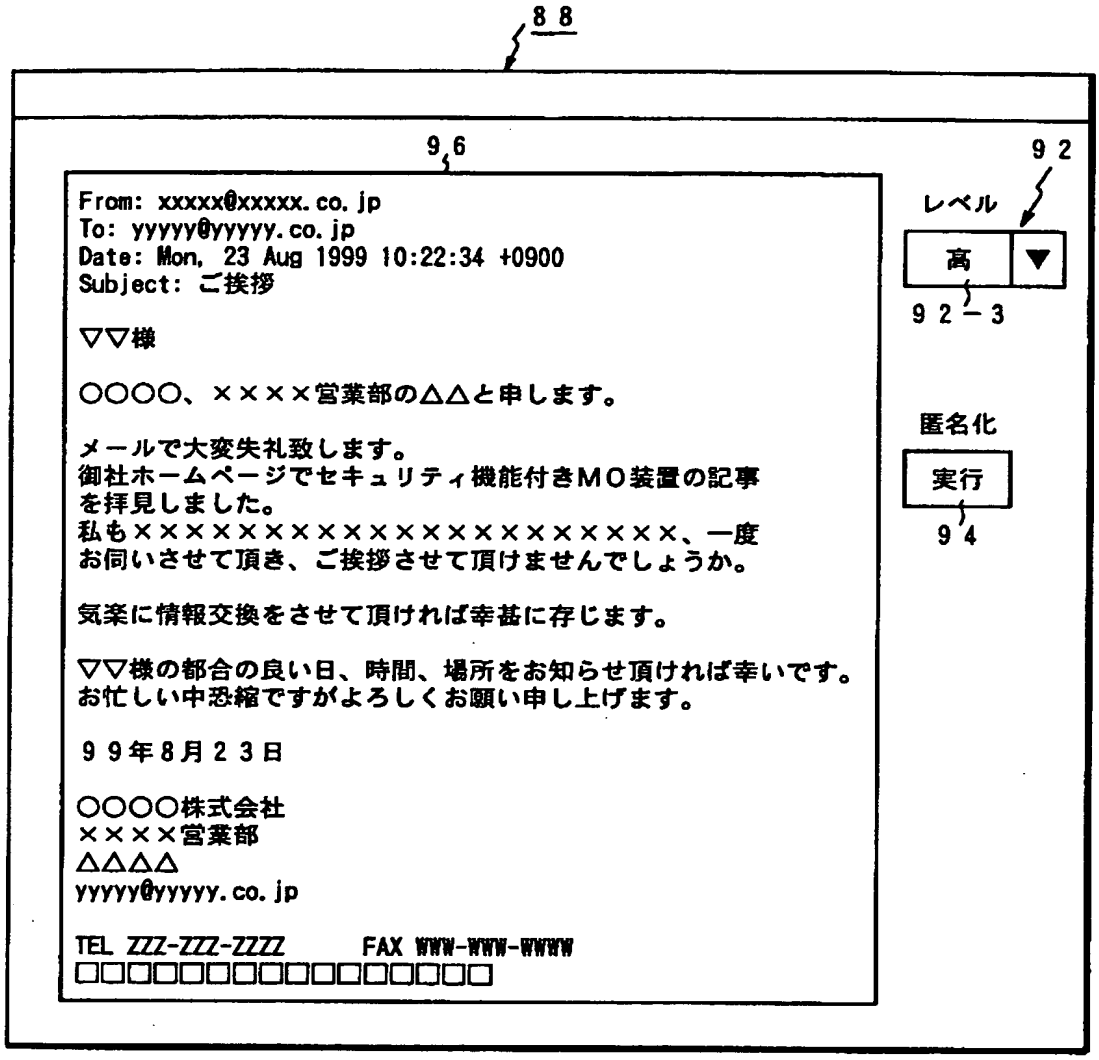
【図１５】

低レベルの閾値を指示した場合の本発明による匿名化文書の画面説明図



【図１６】

高レベルの閾値を指示した場合の本発明による匿名化文書の画面説明図



【書類名】　　要約書

【要約】

【課題】個人情報の隠蔽化を機械化して作業コストを低減し、必要応じて隠蔽化の度合を調整可能とする。

【解決手段】文書匿名化装置は文書を入力する文書入力部１０と、入力文書から匿名対象表記を抽出し、抽出した匿名対象表記がどの程度の強さで個人を特定できるかを評価する特定度を算出する特定度計算部１２と、所定の閾値より大きい特定度を持つ入力文書中の表記を匿名化する匿名化処理部１８とを備える。特定度計算部１８は、入力文書から人名と周辺表記を抽出し、抽出した人名と周辺表記がどの程度の強さで個人を特定できるかを評価する特定度を算出し、匿名化処理部１８は、所定の閾値よりも大きい特定度をもつ人名と周辺表記を、伏せ字化、一般化、低特程度化、暗号化等により匿名化する。

【選択図】　　図１

出　願　人　履　歴　情　報

識別番号　　〔０００００５２２３〕

１．変更年月日　　１９９６年　３月２６日
［変更理由］　住所変更
住　所　　神奈川県川崎市中原区上小田中４丁目１番１号
氏　名　　富士通株式会社